INFLUENCIA DE VARIABLES SOCIALES, ECONÓMICAS Y ESPACIALES EN ENFERMEDADES TRANSMITIDAS POR VECTORES USANDO ALGORITMOS Y TÉCNICAS DE MACHINE LEARNING

JORGE ALBERTO CARDONA GALLEGO

Tesis de Grado para optar al Título de

Máster en Ingeniería de Sistemas y Computación

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERIAS

MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACION

PEREIRA

2018

INFLUENCIA DE VARIABLES SOCIALES, ECONÓMICAS Y ESPACIALES EN ENFERMEDADES TRANSMITIDAS POR VECTORES USANDO ALGORITMOS Y TÉCNICAS DE MACHINE LEARNING

JORGE ALBERTO CARDONA GALLEGO

Director: Rafael Ricardo Rentería Ramos, Post-Ph. D.

Asesor metodológico: Guillermo Roberto Solarte Martínez, Ph. D.

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERIAS

MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACION

PEREIRA

2018

Nota de aceptación

Firma del jurado

Firma del jurado

Pereira, Febrero de 2018

Agradecimiento

Agradezco primero a Dios por permitirme llegar hasta culminar esta maestría y seguidamente a mi mamá y mi abuelita por hacer posible que siempre tuviera el apoyo en todos los sentidos para asistir y rendir en la universidad con todas mis facultades potenciadas.

Al doctor Rafael Ricardo Rentería Ramos, por su ayuda como persona y el aporte absoluto en todos los asuntos relacionados con este trabajo, en especial ahora que fue mi director de tesis.

Por último, a Angela Julieth Gil Gil y a Diana Carolina Acevedo Ramírez quienes además de ser mis colegas y mi familia, me ofrecieron su apoyo incondicional en esta fase de mi existencia, brindándome total convicción para concluir mi primer título de posgrado y por alentarme a continuar mi siguiente meta académica el doctorado.

Dedicatoria

Dedico esta tesis a mi abuelita al igual que a mi mama, seres fantásticos que continuamente me han apoyado en todos mis retos, quienes me brindaron una gran crianza y siempre inculcaron en mí, el respeto, los valores como persona, y la importancia de la academia como propuesta de progreso y crecimiento en la vida.

TABLA DE CONTENIDO

RESUMEN 15
1. PROBLEMA DE INVESTIGACIÓN 18
I. DESCRIPCIÓN DEL PROBLEMA 18
II. FORMULACIÓN DEL PROBLEMA 21
III. OBJETIVOS 22
IV. JUSTIFICACIÓN 23
2. MARCO TEÓRICO 25
I. VIRUS 25
II. VECTORES 28
III. ENFERMEDADES TRANSMITIDAS POR VECTORES 29
IV. VIRUS TRANSPORTADOS POR VECTORES 31
V. DENGUE 32
VI. CASOS DE DENGUE 36
VII. MINERÍA DE DATOS 38
VIII. MACHINE LEARNING 40
3. ESTADO DEL ARTE 42
I. EVOLUCIÓN COLOMBIANA DE LOS VECTORES 42
II. ALGORITMOS DE MACHINE LEARNING 48
III. INVESTIGACIONES DE REFERENCIA 52
4. EXPLORACIÓN DE LA INFORMACIÓN 84
I. ORIGEN DE LOS DATOS 84
II. PREPARACIÓN DE LOS DATOS 84
III. REVISIÓN INICIAL DE LOS DATOS 85
IV. DEPURACIÓN DE LA INFORMACIÓN 85
V. ANÁLISIS PRELIMINAR DE LA INFORMACIÓN 92
VI. CARACTERÍSTICAS DE LA INFORMACIÓN 103
VII. EVALUACIÓN PRELIMINAR DE LAS CARACTERÍSTICAS DE LA INFORMACIÓN 111
5. ANÁLISIS DE LA INFORMACIÓN 114
I. ANÁLISIS RELACIONAL DEL SET DE DATOS 114

II. ELECCIÓN DE CLASE PARA EL ANÁLISIS DE DATOS CON MACHINE LEARNING 123
III. SELECCIÓN DEL MÉTODO DE VALIDACIÓN EN ALGORITMOS DE MACHINE LEARNING 123
IV. ALGORITMOS DE MACHINE LEARNING PARA EL SET DE DATOS 125
V. EVALUACIÓN DE LOS ALGORITMOS DE APRENDIZAJE 133
VI. ALGORITMO DE IDENTIFICACIÓN DE CARACTERÍSTICAS PARA EL SET DE DATOS 134
VII. ALGORITMOS FACTIBLES PARA EL SET DE DATOS 142
VIII. ALGORITMOS ELEGIDOS PARA ANÁLISIS EXPLICATIVO DEL SET DE DATOS 147
IX. EVALUACIÓN DE VARIABLES EXPLICATIVAS DEL MODELO 148
MÉTODO STEPWISE 149
X. ZONAS A EVALUAR EN EL SET DE DATOS 175
XI. NIVEL DE PREDICCIÓN DE LOS CASOS DE DENGUE 175
6. CONCLUSIONES 183
7. RECOMENDACIONES 188
8. REFERENCIAS 189

ÍNDICE DE TABLAS

Tabla 1.Tipos de virus según el Grupo..... 26
Tabla 2. Tipos de enfermedades de transmisión vertical. 27
Tabla 3. Tipos de enfermedades de transmisión horizontal. 28
Tabla 4. Actores en la Cadena Epidemiológica..... 29
Tabla 5. Vectores del Flavivirus. 30
Tabla 6. Vectores del PestivirusÉ 30
Tabla 7. Vectores del Hepacivirus..... 30
Tabla 8. Enfermedades transmitidas por mosquitos. 31
Tabla 9. Casos de dengue en Colombia 2008-2017. 36
Tabla 10. Comparación Dengue en Colombia contra el continente americano..... 37
Tabla 11. Comparación Muertes por Dengue en Colombia contra el continente americano. 37
Tabla 12. Símbolos Municipios. 87
Tabla 13. Comunas por municipios 1. 88
Tabla 14. Comunas por municipios 2. 89
Tabla 15. Símbolos de etnias. 90
Tabla 16. Símbolos de estratos. 90
Tabla 17. Símbolos de seguridad social. 90
Tabla 18. Símbolos de grupo etario. 91
Tabla 19. Símbolos de Tipo de Caso. 91
Tabla 20. Símbolos de tipos de área de ocupación. 91
Tabla 21. Símbolos de profesiones. 92
Tabla 22. Algoritmos Seleccionados para evaluación del set de datos. 126
Tabla 23. Evaluación del rendimiento del estimador del algoritmo anual set de datos completo. 133
Tabla 24. Evaluación del rendimiento del estimador del algoritmo anual set de datos confirmados. 133
Tabla 25. Algoritmos de Identificación de Características. 135
Tabla 26. Características Clases Principales Datos Completos. 141
Tabla 27. Características Clases Principales Datos Confirmado. 141
Tabla 28. Comparación anual rendimiento algoritmos seleccionados datos completos. 146
Tabla 29. Comparación anual rendimiento algoritmos seleccionados datos confirmados. 146
Tabla 30. Resumen Importancia porcentual Clases del modelo. 149
Tabla 31. Comparación anual rendimiento algoritmos seleccionados sin clase COMUNA datos completos. 151
Tabla 32. Comparación anual rendimiento algoritmos seleccionados sin clase COMUNA datos confirmados. 151

Tabla 33. Comparación anual rendimiento algoritmos seleccionados sin clase ESTRATO datos completos. 152
Tabla 34. Comparación anual rendimiento algoritmos seleccionados sin clase ESTRATO datos confirmados. 152
Tabla 35. Comparación anual rendimiento algoritmos seleccionados sin clase OCUPACIÓN datos completos. 153
Tabla 36. Comparación anual rendimiento algoritmos seleccionados sin clase OCUPACIÓN datos confirmados. 153
Tabla 37. Comparación anual Eliminación Experimental ETNIA datos completos. 154
Tabla 38. Comparación anual Eliminación Experimental ETNIA datos confirmados. . 154
Tabla 39. Comparación anual rendimiento algoritmos seleccionados sin clase GRUPO ETARIO datos completos. 155
Tabla 40. Comparación anual rendimiento algoritmos seleccionados sin clase GRUPO ETARIO datos confirmados. 155
Tabla 41. Comparación anual rendimiento algoritmos seleccionados sin clase ÁREA datos completos. 156
Tabla 42. Comparación anual rendimiento algoritmos seleccionados sin clase ÁREA datos confirmados. 156
Tabla 43. Comparación anual rendimiento algoritmos seleccionados sin clase SEGURIDAD SOCIAL datos completos. 157
Tabla 44. Comparación anual rendimiento algoritmos seleccionados sin clase SEGURIDAD SOCIAL datos confirmados. 157
Tabla 45. Comparación anual rendimiento algoritmos seleccionados sin clase ETNIA datos completos. 159
Tabla 46. Comparación anual rendimiento algoritmos seleccionados sin clase ETNIA datos confrmados. 159
Tabla 47. Método Backward sin clase ETNIA – ÁREA datos completos. 160
Tabla 48. Método Backward sin clase ETNIA – ÁREA datos confirmados. 160
Tabla 49. Método Backward sin clase ETNIA - ÁREA – ETARIO datos completos. ... 161
Tabla 50. Método Backward sin clase ETNIA - ÁREA – ETARIO datos confirmados. 161
Tabla 51. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS datos completos. 162
Tabla 52. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS datos confirmados. 162
Tabla 53. Método Backward sin clase ETNIA - ÁREA - ETARIO - SS – OCUPACIÓN datos completos. 163
Tabla 54. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS – OCUPACIÓN datos confirmados. 163
Tabla 55. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN – ESTRATO datos completos. 164
Tabla 56. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN – ESTRATO datos confirmados. 164

Tabla 57. Método Mayor-Menor sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO – COMUNA datos completos. 165
Tabla 58. Método Mayor-Menor sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO – COMUNA datos confirmados. 165
Tabla 59. Método Mayor-Menor sin clase COMUNA datos completos. 168
Tabla 60. Método Mayor-Menor sin clase COMUNA datos confirmados. 168
Tabla 61. Método Mayor-Menor sin clase COMUNA - ESTRATO datos completos. 169
Tabla 62. Método Mayor-Menor sin clase COMUNA - ESTRATO datos confirmados. 169
Tabla 63. Método Mayor-Menor sin clase COMUNA – ESTRATO - OCUPACIÓN datos completos. 170
Tabla 64. Método Mayor-Menor sin clase COMUNA - ESTRATO - OCUPACIÓN datos confirmados. 170
Tabla 65. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN - SS datos completos. 171
Tabla 66. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN - SS datos confirmados. 171
Tabla 67. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN – SS - ETARIO datos completos. 172
Tabla 68. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN – SS - ETARIO datos confirmados. 172
Tabla 69. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO - ÁREA datos completos. 173
Tabla 70. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO - ÁREA datos confirmados. 173
Tabla 71. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA - ETNIA datos completos. 174
Tabla 72. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO - ÁREA - ETNIA datos confirmados. 174
Tabla 73. Capacidad Predictiva datos Completos. 176
Tabla 74. Capacidad Predictiva datos Depurados. 177
Tabla 75. Capacidad Predictiva Pereira datos Completos. 178
Tabla 76. Capacidad Predictiva Pereira datos Depurados. 178
Tabla 77. Capacidad Predictiva Dosquebradas datos Completos. 179
Tabla 78. Capacidad Predictiva Dosquebradas datos Depurados. 179
Tabla 79. Capacidad Predictiva La Virginia datos Completos. 180
Tabla 80. Capacidad Predictiva La Virginia datos Depurados. 180
Tabla 81. Capacidad Predictiva Santa Rosa datos Completos. 181
Tabla 82. Capacidad Predictiva Santa Rosa datos Depurados. 181
Tabla 83. Resumen Aprendizaje Anual DT vs Ciudad Datos Completos. 182
Tabla 84. Resumen Aprendizaje Anual DT vs Ciudad Datos Depurados. 182

ÍNDICE DE ILUSTRACIONES

Ilustración 1. Distribución serotipos de virus de dengue en las Américas 1990-2014... 34
Ilustración 2. Incidencia del dengue en las Américas 1980-2014..... 34
Ilustración 3. Casos Totales Reportados de dengue entre 2008 y 2014. 93
Ilustración 4. Casos Confirmados Reportados de dengue entre 2008 y 2014. 94
Ilustración 5. Casos Totales Por Municipio. 94
Ilustración 6. Casos Confirmados por Municipio. 95
Ilustración 7. Casos Completos Anuales por Municipio..... 96
Ilustración 8. Casos Confirmados Anuales por Municipio. 97
Ilustración 9. Casos totales por Municipio y Año con mayor cantidad de casos 2008 - 2014. 97
Ilustración 10. Casos totales por Municipio y Año con mayor cantidad de casos 2008 - 2009-2011-2012-2013-2014..... 98
Ilustración 11. Casos Totales Anuales por Comunas de Pereira. 99
Ilustración 12. Casos Confirmados Anuales por Comunas de Pereira..... 100
Ilustración 13. Casos Totales Anuales por Comunas de Dosquebradas..... 101
Ilustración 14. Casos Confirmados Anuales por Comunas de Dosquebradas. 101
Ilustración 15. Casos Totales Anuales por Comunas de Santa Rosa. 102
Ilustración 16. Casos Confirmados Anuales por Comunas de Santa Rosa..... 102
Ilustración 17. Casos globales por género y set de datos. 103
Ilustración 18. Casos Anuales Género Casos Completa..... 104
Ilustración 19. Casos Anuales Género Casos Confirmados..... 104
Ilustración 20. Tipos de áreas donde se confirmaron los casos. 104
Ilustración 21. Cantidad de Casos por Área Casos Completas..... 105
Ilustración 22. Cantidad de Casos por Área Casos Confirmados. 105
Ilustración 23. Casos por grupo etario anuales. 105
Ilustración 24. Grupos Etarios Datos Completos..... 106
Ilustración 25. Grupos Etarios Datos Confirmados. 106
Ilustración 26. Casos de Tipos de estratos por año. 106
Ilustración 27. Casos de dengue por estrato, datos Completos. 107
Ilustración 28. Casos de dengue por estrato, datos Confirmados..... 107
Ilustración 29. Tipos de seguridad social por año donde se confirmaron los casos. ... 108
Ilustración 30. Casos de Seguridad Social datos Completos. 108
Ilustración 31. Casos de Seguridad Social datos Confirmados..... 108
Ilustración 32. Etnias que reportan casos por año. 109
Ilustración 33. Casos anuales por Etnia casos Completos..... 109
Ilustración 34. Casos anuales por Etnia casos Confirmados. 109
Ilustración 35. Casos ocupaciones datos Completos..... 110
Ilustración 36. Casos ocupaciones datos Confirmados..... 110
Ilustración 37. Análisis de Correspondencia Municipio vs Género..... 116
Ilustración 38. Análisis de Correspondencia Municipio vs Tipos de Área..... 116

Ilustración 39. Análisis dimensional Área vs Municipio. 117
Ilustración 40. Análisis de Correspondencia Municipio vs Grupo Etario..... 118
Ilustración 41. Análisis dimensional Municipio vs Grupo Etario..... 118
Ilustración 42. Análisis de Correspondencia Municipio vs Estratos..... 119
Ilustración 43. Análisis dimensional Municipio vs Estratos..... 119
Ilustración 44. Análisis de Correspondencia Municipio vs Seguridad Social..... 120
Ilustración 45. Análisis dimensional Municipio vs Seguridad Social..... 120
Ilustración 46. Análisis de Correspondencia Municipio vs Etnias..... 121
Ilustración 47. Análisis dimensional Municipio vs Etnias..... 121
Ilustración 48. Análisis de Correspondencia Municipio vs Ocupación..... 122
Ilustración 49. Análisis dimensional Municipio vs Ocupación..... 122
Ilustración 50. Rendimiento algoritmo SVMC. Ambos sets de datos. 127
Ilustración 51. Rendimiento algoritmo SVMR. todos los datos..... 127
Ilustración 52. Rendimiento algoritmo LM. Ambos sets de datos..... 128
Ilustración 53. Rendimiento algoritmo LR. Ambos sets de datos. 128
Ilustración 54. Rendimiento algoritmo MLPC. Ambos sets de datos..... 129
Ilustración 55. Rendimiento algoritmo MLPR. Ambos sets de datos..... 129
Ilustración 56. Rendimiento algoritmo GBC. Ambos sets de datos. 130
Ilustración 57. Rendimiento algoritmo LASSO. Ambos sets de datos..... 130
Ilustración 58. Rendimiento algoritmo DT. Ambos sets de datos. 131
Ilustración 59. Rendimiento algoritmo RF. Ambos sets de datos. 131
Ilustración 60. Comparación algoritmos set de datos completo. 132
Ilustración 61. Comparación algoritmos set de datos confirmados. 132
Ilustración 62. Características Principales RFC Datos Completos..... 135
Ilustración 63. Características Principales RFC Datos Confirmados..... 136
Ilustración 64. Características Principales RFR Datos Completos..... 136
Ilustración 65. Características Principales RFR Datos Confirmados..... 137
Ilustración 66. Características Principales ETC Datos Completos. 137
Ilustración 67.. Características Principales ETC Datos Confirmados..... 138
Ilustración 68. Características Principales ETR Datos Completos. 138
Ilustración 69. Características Principales ETR Datos Confirmados..... 139
Ilustración 70. Características Principales ABC Datos Completos..... 139
Ilustración 71. Características Principales ABC Datos Confirmados..... 140
Ilustración 72. Características Principales ABR Datos Completos..... 140
Ilustración 73. Características Principales ABR Datos Confirmados..... 141
Ilustración 74. Evaluación Algoritmo MLPR Sin la clase MES. 143
Ilustración 75. Evaluación Algoritmo LASSO Sin la clase MES. 143
Ilustración 76. Evaluación Algoritmo RF Sin la clase MES..... 144
Ilustración 77. Evaluación Algoritmo DT Sin la clase MES..... 144
Ilustración 78. Resumen Algoritmos Sin la clase MES Set de Datos Completo..... 145
Ilustración 79. Resumen Algoritmos Sin la clase MES Set de Datos Confirmados..... 145
Ilustración 80. Eliminación Experimental DT COMUNA. 150
Ilustración 81. Eliminación Experimental RF COMUNA. 150

Ilustración 82. Eliminación Experimental RF ESTRATO. 151
Ilustración 83. Eliminación Experimental DT ESTRATO. 151
Ilustración 84. Eliminación Experimental DT OCUPACIÓN. 152
Ilustración 85. Eliminación Experimental RF OCUPACIÓN. 152
Ilustración 86. Eliminación Experimental RF ETNIA..... 153
Ilustración 87. Eliminación Experimental DT ETNIA..... 153
Ilustración 88. Eliminación Experimental RF GRUPO ETARIO..... 154
Ilustración 89. Eliminación Experimental DT GRUPO ETARIO..... 154
Ilustración 90. Eliminación Experimental DT ÁREA. 155
Ilustración 91. Eliminación Experimental RF ÁREA. 155
Ilustración 92. Eliminación Experimental DT SEGURIDAD SOCIAL..... 156
Ilustración 93. Eliminación Experimental RF SEGURIDAD SOCIAL..... 156
Ilustración 94. Método Backward RF ETNIA. 158
Ilustración 95. Método Backward DT ETNIA. 158
Ilustración 96. Método Backward DT ETNIA – ÁREA. 159
Ilustración 97. Método Backward RF ETNIA – ÁREA. 159
Ilustración 98. Método Backward RF ETNIA - ÁREA - ETARIO..... 160
Ilustración 99. Método Backward DT ETNIA - ÁREA – ETARIO..... 160
Ilustración 100. Método Backward DT ETNIA - ÁREA - ETARIO - SS..... 161
Ilustración 101. Método Backward RF ETNIA - ÁREA - ETARIO - SS..... 161
Ilustración 102. Método Backward RF ETNIA - ÁREA - ETARIO - SS - OCUPACIÓN. 162
Ilustración 103. Método Backward DT ETNIA - ÁREA - ETARIO - SS - OCUPACIÓN. 162
Ilustración 104. Método Backward RF ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO. 163
Ilustración 105. Método Backward DT ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO. 163
Ilustración 106. Método Mayor-Menor RF ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN – ESTRATO - COMUNA..... 164
Ilustración 107. Método Mayor-Menor DT ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO - COMUNA. 164
Ilustración 108. Método Mayor-Menor DT COMUNA. 167
Ilustración 109. Método Mayor-Menor RF COMUNA. 167
Ilustración 110. Método Mayor-Menor RF COMUNA - ESTRATO. 168
Ilustración 111. Método Mayor-Menor DT COMUNA - ESTRATO. 168
Ilustración 112. Método Mayor-Menor RF COMUNA - ESTRATO - OCUPACIÓN. 169
Ilustración 113. Método Mayor-Menor DT COMUNA - ESTRATO - OCUPACIÓN. 169
Ilustración 114. Método Mayor-Menor DT COMUNA – ESTRATO – OCUPACIÓN - SS. 170
Ilustración 115. Método Mayor-Menor RF COMUNA – ESTRATO – OCUPACIÓN - SS. 170

Ilustración 116. Método Mayor-Menor RF COMUNA – ESTRATO – OCUPACIÓN – SS - ETARIO..... 171
Ilustración 117. Método Mayor-Menor DT COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO..... 171
Ilustración 118. Método Mayor-Menor DT COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA. 172
Ilustración 119. Método Mayor-Menor RF COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA. 172
Ilustración 120. Método Mayor-Menor RF COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA – ETNIA..... 173
Ilustración 121. Método Mayor-Menor DT COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA - ETNIA. 173
Ilustración 122. Capacidad de Aprendizaje DT Todos los datos. 176
Ilustración 123. Capacidad de Aprendizaje RF Todos los datos. 176
Ilustración 124. Capacidad de Predicción RF Pereira..... 178
Ilustración 125. Capacidad de Predicción DT Pereira. 178
Ilustración 126. Capacidad de Predicción DT Dosquebradas. 179
Ilustración 127. Capacidad de Predicción RF Dosquebradas. 179
Ilustración 128. Capacidad de Predicción DT La Virginia..... 180
Ilustración 129. Capacidad de Predicción RF La Virginia..... 180
Ilustración 130. Capacidad de Predicción DT Santa Rosa. 181
Ilustración 131. Capacidad de Predicción RF Santa Rosa. 181

RESUMEN

Colombia es un país donde las enfermedades trasmitidas por vectores como el dengue continúan siendo un problema de salud pública a tener en cuenta, el vector continuamente ha logrado colonizar y expandirse a gran cantidad de zonas en todo el territorio nacional venciendo múltiples tipos de barreras para conseguirlo, haciendo de esto un factor recurrente y con una mayor incidencia a través de los años. Aportar un estudio donde se propongan que factores como los sociales, económicos y espaciales pueden favorecer el crecimiento, permanencia e incidencia del flagelo en una población, contribuyen adicionalmente en el entendimiento de cómo se fortalece el fenómeno, ayudando en la construcción de modelos explicativos para la obtención de pronósticos precisos y oportunos de la incidencia del dengue en Colombia y puntualmente en el departamento de Risaralda.

Palabras clave — Aedes, Algoritmos, Correlación, Clasificadores, Dengue, Enfermedad, Epidemia, Extracción de Características, Machine Learning, Minería de datos, OMS, OPS, Probabilidad, Risaralda, Set de datos, Sivigila, Variables Sociales, Económicas y Espaciales, Vectores, Virus.

APORTES

Entre los beneficios que este estudio puede aportar a la comunidad científica, académica y a los tomadores de decisiones en política pública y de salud están:

Determinar la importancia que generan las variables del objetivo de este estudio sobre una región en enfermedades transmitidas por vectores, para que futuras investigaciones sobre enfermedades de carácter masivo, tengan en cuenta estas características logrando una mayor perspectiva de la incidencia de un fenómeno, permitiendo crear un modelo evaluativo más amplio que podría explicar de mejor manera un problema dado.

Tomar en cuenta el impacto de los factores económicos, sociales y espaciales de una región, para los planes de gobierno en política pública y proyectos de salud para evitar y disminuir la propagación e incremento del riesgo de vectores epidemiológicos.

Disminuir los costos que puede acarrear la atención de una epidemia de virus dengue, promoviendo campañas sociales oportunas en zonas identificadas como vulnerables al ataque del vector.

Proponer un modelo de machine learning que permita evaluar de manera eficiente el caso de estudio, permitiendo desarrollar cálculos efectivos de los casos de dengue en una zona, para así lograr anticiparse a un posible foco de infección.

INTRODUCCIÓN

La información de un fenómeno a través de los años se ha convertido en la fuente principal del avance en el campo de la ciencia a la que pertenezca. Con el paso del tiempo han surgido métodos y técnicas que permiten evaluar dichos datos de manera profunda y detallada logrando encontrar datos importantes relacionados con el caso de estudio, estas técnicas en la actualidad se basan en campos del conocimiento muy específicos como por ejemplo la inteligencia artificial, que dentro de esta tiene unos campos interesantes que estudian la información, y que siguen evolucionando constantemente hasta llegar a ser necesaria una rama exclusivamente para esta tarea como la ciencia de datos.

Las organizaciones y equipos de investigación actualmente están muy interesadas en descubrir los patrones que se presentan entre algún evento y los directamente implicados con este, para analizar ese tipo de información se puede usar un campo de la inteligencia artificial que se ajusta a este modelo tales como el machine learning, el cual aprende de la información que tiene a su disposición y posteriormente logra predecir sucesos que podrían ocurrir relacionados con los datos que ya aprendió o con información nueva que tenga relación sobre el caso de estudio. Entonces una buena opción para esta tesis, es usar un campo de la inteligencia artificial que permita analizar la información de este trabajo, para tratar de proponer un modelo explicativo del fenómeno, una de las maneras viables que permiten cumplir con este objetivo, es utilizando algoritmos y técnicas de machine learning.

1. PROBLEMA DE INVESTIGACIÓN

I. DESCRIPCIÓN DEL PROBLEMA

Según el Instituto Geográfico “Agustín Codazzi”, Colombia cuenta con seis regiones naturales y están definidas como región amazónica, región andina, región caribe, región insular, región Orinoquía, región pacífica (IGAC, Instituto Geografico Agustín Codazzy, 2002), cada uno de estos territorios cuenta con condiciones geográficas y demográficas diferentes, esto posibilita que la población de una región tienda a ser afectada o no en mayor proporción por algún tipo de vulnerabilidad, como por ejemplo en el campo de la salud.

Conforme plantea la Comisión Económica para América Latina y el Caribe “Los problemas ambientales conciernen a las ciencias exactas, naturales y sociales. Pero al mismo tiempo son problemas que involucran decisiones políticas, a veces controvertidas y por lo mismo muy difíciles de resolver” (de la Fuente, 2017), este tipo de situaciones hacen que las condiciones naturales de un sitio sean forzadas en gran cantidad por el desarrollo de las regiones y el interés económico, generando cambios ambientales que repercuten directamente en el entorno donde se encuentren, modificando las circunstancias de los pobladores y su nivel de fragilidad, como menciona el Ministerio de Ambiente y Desarrollo Sostenible de Colombia, en su publicación Diagnóstico Nacional De Salud Ambiental, “La estrecha relación entre la salud humana y el ambiente se manifiesta más frecuentemente en poblaciones que por características particulares, se encuentran en una situación de vulnerabilidad o susceptibilidad” (Ministerio de Ambiente, 2012). Adicionando a esto que en varios poblados de Colombia por temas como el orden

público asociado al conflicto armado o por situaciones como el tráfico de drogas hacen que el medio ambiente tenga un deterioro mayor en zonas de conflicto, tal como se propone en la publicación de Martínez y Londoño “El medio ambiente es un actor implícito o pasivo en la lucha de poderes; sin embargo, directamente afectado en la misma.” (Martínez y Londoño, 2017). Por esto es más difícil subsanar estos problemas y los factores de riesgo sobre los pobladores suelen aumentar.

Los vectores son organismos que transmiten patógenos, entre ellos parásitos, de una persona (o animal) infectada a otra y ocasionan enfermedades graves en el ser humano como el dengue, el zika y chikungunya. Estas enfermedades son más frecuentes en zonas tropicales y subtropicales, y en lugares con problemas de acceso al agua potable y al saneamiento. Las enfermedades vectoriales representan un 17% de la carga mundial estimada de enfermedades infecciosas. La más mortífera de todas ellas (el paludismo) causó 627.000 muertes en 2012; no obstante, la enfermedad de este tipo con mayor crecimiento en el mundo es el dengue, cuya incidencia se ha multiplicado por 30 en los últimos 50 años (OMS, Campañas mundiales de salud pública de la OMS, 2017). Es necesario considerar la movilidad de población en el país hacia regiones que reciben muchas personas en su zona como por ejemplo el eje cafetero, esto permite la posibilidad de que el vector se polarice en las zonas donde los infectados lleguen y genere un riesgo potencial de infección sobre los nativos de las regiones que están libres de contagio.

Hay poblaciones que tienden a padecer en mayor medida un tipo de afección en comparación con otras regiones, causadas por agentes externos como los mosquitos

Aedes aegypti, vectores que se encuentran a alturas menores a los 2200 metros sobre el nivel del mar; es importante resaltar que 951 de los 1.096 municipios de Colombia están en ese rango (IGAC, Notas Geográficas, 2017); siendo zonas potenciales de influencia para la acción de los vectores, lo cual representa que el 86.77% del territorio nacional cuenta con las condiciones medio ambientales necesarias para que el vector se establezca. Un factor que los agentes infecciosos aprovechan para su expansión e intensificación es que existen zonas que debido a su ubicación geográfica de difícil acceso o que se encuentran bajo el control de grupos ilegales hacen que se disminuya, se retrase u omita la distribución o recepción de ayudas en ocasiones de carácter humanitario que buscan mejorar la calidad de vida a los residentes (CRUZ ROJA, 2017), esto obstaculiza que el control sobre los sujetos contagiados, los ignotos y los vectores sea más efectivo, posibilitando el aumento de casos positivos.

Las enfermedades más comunes transmitidas por vectores son el dengue, esquitosomiasis, filariasis linfática, fiebre hemorrágica, fiebre amarilla, changas, paludismo, tripanosomiasis, zika y chikungunya. La propagación de enfermedades transmitidas por vectores puede estar dadas por Los cambios en las prácticas agrícolas debido a las variaciones de temperatura y precipitaciones pueden influir en la propagación de enfermedades transmitidas por vectores. La información climática se puede utilizar para vigilar y predecir a largo plazo la distribución y las tendencias del paludismo y otras enfermedades variables en función del clima; el cambio de comportamiento es un elemento crucial en lo concerniente a las enfermedades transmitidas por vectores. La OMS colabora con asociados a fin de difundir

conocimientos y mejorar la sensibilización, de manera que las personas sepan cómo protegerse a sí mismas y proteger a sus comunidades contra mosquitos, garrapatas, chinches, moscas y otros vectores (OMS, Enfermedades transmitidas por vectores, 2017).

Colombia cuenta con diversidad cultural especialmente delimitada por las fronteras naturales que permiten que el impacto de dichos vectores sea más voraz sobre algunas localidades ya que la superficie del país posee condiciones biofísicas, climáticas y sociales, aptas para la propagación de la enfermedad trasmitida por vectores, ya que estas son de tipo endémico (Ministerio de Salud y. P., 2017),“El dengue en Colombia representa un problema prioritario en salud pública debido a la re emergencia e intensa transmisión con tendencia creciente, el comportamiento de ciclos epidémicos cada dos o tres años, el aumento en la frecuencia de brotes de dengue hemorrágico y síndrome de choque por dengue, la circulación simultánea de diferentes serotipos, la reintroducción del serotipo tres, la infestación por aedes aegypti de más de 90% del territorio nacional situado por debajo de los 2.200 msnm(metros sobre el nivel del mar), la introducción de Aedes albopictus y la urbanización de la población por problemas de violencia” (Grupo de vigilancia y control, 2010).

II. FORMULACIÓN DEL PROBLEMA

¿Cómo caracterizar los diferentes tipos de variables que influencian el estado epidemiológico de una población o al interior de un territorio?

III. OBJETIVOS

OBJETIVO GENERAL

Valorar la influencia de variables sociales, económicas y espaciales en enfermedades transmitidas por vectores usando algoritmos y técnicas de machine learning.

OBJETIVOS ESPECÍFICOS

Conceptualizar el marco teórico – metodológico para el análisis de las enfermedades transmitidas por vectores y su influencia económica y social.

Identificar patrones espaciales, económicos y sociales que inciden en el favorecimiento de la población de vectores y el aumento de casos en zonas vulnerables, usando el estudio de datos para determinar los agentes coincidentes que más inciden en la proliferación del virus.

Analizar los algoritmos de evaluación de datos para definir qué tipo de algoritmo computacional favorece una mejor identificación de los factores que promueven la aparición de los vectores, según el caso de estudio requerido.

Evaluar la información del sivigila relacionada con el evento de tipo dengue entre los años 2008 a 2014 pertenecientes al departamento de Risaralda con algoritmos y técnicas de machine learning.

Definir si las variables sociales, económicas y espaciales de una región favorecen el establecimiento, propagación y crecimiento del virus sobre las zonas que presentan brotes y ataques por parte de este tipo de vectores y ordenar la información sobre cada una de ellas.

IV. JUSTIFICACIÓN

Los problemas de salud pública en las naciones son un tema de gran interés para los organismos de vigilancia y control estatal, todo país puede sufrir un episodio desafortunado que conlleve a grandes pérdidas tanto humanas como económicas, de ahí la importancia de poder contrarrestar este tipo de flagelos con técnicas que permitan identificar situaciones que se estén dando en un sector o que se puedan llegar a dar.

Con el aporte de técnicas o herramientas que permitan analizar problemas de este tipo, es posible minimizar el impacto sobre la población y la economía de ser posible en muchos casos prevenirlos y evitar una catástrofe, por eso se hace necesario contar con este tipo de soluciones para monitorear, analizar y actuar sobre una situación dada.

Este proyecto cuenta con una pertinencia social con el que se podrán evidenciar las zonas que tienen población vulnerable al ataque de los vectores, esto ayudaría a entender y a conocer la calidad de vida de vida de los pobladores en zonas tanto infectadas, como en zonas con potencial de ataque para los mosquitos, este tipo de información permite identificar patrones característicos de propagación del virus y lograr identificar criterios evaluativos antes de que suceda la enfermedad que indique que los

vectores se van a sentar en ese lugar y sea posible generar una mejor respuesta, permitiendo una mejor distribución de los recursos para tratar esas poblaciones, y poder generar programas que sean propios para la atención y la disminución de variables que afectan a cada población. También se puede favorecer sobre la parte económica de cada región, ya que la inversión pública que tienen que hacer las administraciones para dar respuesta y atender este tipo de pandemias, es elevado y debe ser inmediato para evitar lo proliferación de casos y así prevenir crisis mayores en las zonas afectadas, con esto se podría generar capacitaciones a las personas que carecen de recursos para que tengan los cuidados básicos que permitan prevenir o multiplicar ese tipo de situaciones y con esto no tengan que intervenir solo en la fase de atención de un tema grave de salud pública.

Como componente adicional se tiene la parte tecnológica, dado que la ayuda que puede dar una implementación de un proceso (herramienta o instrumento) que permita evaluar y divisar la situación de una región o población se convierte en un fuerte aporte para los tomadores de decisiones en el país con los cuales se pueden apoyar y entender de manera más clara la actualidad y efectos del vector en un sector dado, brindando la oportunidad de realizar simulaciones de posibles situaciones, evaluación de casos y estudios anteriores. Se puede agregar a favor que la solución puede ser escalable, además se desarrolla con herramientas libres. Finalmente, los estudios y resultados pueden ser replicados de manera idéntica que la investigación original.

2. MARCO TEÓRICO

I. VIRUS

Los virus como agentes infecciosos en los seres vivos, especialmente en el hombre, han ganado terreno dentro de la historia debido a su dominio sobre las poblaciones en las que tuvo o tiene bajo su influencia y esto básicamente por sus repercusiones al causar gran poder de devastación en la mayoría de los casos sobre los organismos que fueron infectados, teniendo como antecedente que esto ha venido ocurriendo desde casi el comienzo de los orígenes de la vida en el planeta.

En los últimos años del siglo XIX la etiología avanzó en varios temas como las enfermedades de tipo infecciosas, aunque faltaban por resolver muchas afecciones en las plantas, animales y el mismo hombre en las que no se lograba determinar qué tipo de organismo causaba esos problemas. Ya para el siglo XX no se pudo encontrar un hongo, un protozoario o una bacteria como la responsable de estos males, de ahí se descubrieron los virus como autores de las enfermedades de tipo infecciosas.

Gracias a los avances de la biología molecular y la microscopia en los finales del siglo XX fue posible identificar y aislar los virus. “Los virus poseen un solo tipo de ácido nucleico de pequeño tamaño con respecto a otros agentes biológicos, rodeado por una cáscara o cápside formada por numerosas copias de una proteína o de un número limitado de ellas” (Higiene, 2006) Los virus se definen como parásitos intracelulares obligados, que dependen de factores celulares para poder completar su ciclo vital (Demirov, 2004). Estos Son considerados patógenos destructores para plantas,

animales, humanos, bacterias, hongos y levaduras (Neuman, 2008) e incluso existen algunos tipos de virus que pueden parasitar la maquinaria ensamblada por otros virus, como por ejemplo es el caso del virus Sputnik, que parasita la factoría viral de los Mimivirus (La Scola, 2008). Los Mimivirus son virus de gran tamaño que ensamblan su factoría viral en el citoplasma de una ameba, Acanthamoeba polyphaga (La Scola B. Z., 2003).

Los virus pueden ser clasificados de acuerdo a diferentes tipos de criterios como su morfología, su función o por la enfermedad que causan. También existe un parámetro conocido como la clasificación de Baltimore el cual esencialmente agrupa los virus por su genoma y usa el acido desoxiribonucleico (ADN) y el acido ribonucleico (ARN) para su distinción en ingles DNA y RNA, esta clasificación se da en siete grupos como se muestra a continuación en la adaptación de la tabla1 (Arshan Nasir, 2017).

Tabla 1.Tipos de virus según el Grupo.

GRUPO	TIPO DE VIRUS
I	dsDNA
II	ssDNA
III	dsRNA
IV	(+) ssRNA
V	(-) ssRNA
VI	RNA-RT
VII	DNA-RT

Dentro del grupo IV de esta clasificación se encuentra la familia del flavivirus, que se traduce del latín amarillo, constituyen uno entre los 3 géneros que forman la familia viral de Flaviviridae, esta hace parte de una suma cercana a los 70 virus envueltos en los que su genoma se encuentra basado por ARN con polaridad positiva como se observa

en la tabla anterior, existe otro género llamado Pestivirus (del latín pestis traducido como plaga) y el ultimo género conocido como Hepacivirus (del griego y que traduce hígado), los dos últimos se comportan de manera parecida en su replicación, así como los del tipo flavivirus, los grupos son antigénicamente diferentes, no se transmiten por artrópodos y representan familias que se alejaron casi desde los inicios de la evolución en esta familia viral. (Calisher C. H., 2003). Históricamente se tiene conocimiento de que el vector padre de todos los pertenecientes a estas 80 especies que corresponden al tipo flavivirus es originario del continente africano (Endy, 2010).

Las enfermedades de tipo infecciosas son consideradas como la irrupción de microorganismos invasores que son de carácter nocivo en un receptor, los invasores solo logran sobrevivir cuando existen las condiciones que favorezcan la transmisión de estos a un receptor susceptible. Para controlar estos agentes infecciosos y evitar su propagación es necesario conocer las maneras en que estos se pueden transmitir. La forma de transmisión de estos microorganismos puede ser de manera horizontal o vertical (Abizanda, 2001) como se muestra en la adaptación de las tablas 2 y 3.

Tabla 2. Tipos de enfermedades de transmisión vertical.

ENFERMEDADES DE TRANSMISIÓN VERTICAL		
PERINATAL	**NEONATAL**	**CONGÉNITA**
Estas se manifiestan desde el momento del nacimiento, y puede ser producida por un trastorno durante el desarrollo del embrión o durante el parto.	Estas se manifiestan posterior al parto y se da por el contacto directo del recién nacido con el patógeno.	Estas son adquiridas desde la etapa del embarazo

Tabla 3. Tipos de enfermedades de transmisión horizontal.

ENFERMEDADES DE TRANSMISIÓN HORIZONTAL	
DIRECTA	**INDIRECTA**
Se da cuando un organismo infectado transmite la enfermedad o infección a un receptor susceptible mediante el contacto físico.	Se da cuando por intermedio de un canal de infección de tipo vivo o inanimado, se transmite la infección entre un organismo infeccioso a otro susceptible.

II. VECTORES

Un vector es un medio que lleva uno o varios microorganismos al interior de receptores y ayuda a su multiplicación y replicación. Como medios de transmisión los vectores deben tener la capacidad de introducirse en la célula receptora, al conseguirlo debe replicarse y generar múltiples copias de sí mismo (Kok, 1984). Los vectores cuentan con una unidad de replicación básica, la selección en la célula receptora puede venir de uno o varios genes que permiten la selección y sitios de restricción que favorecen la adición del ADN exógeno (Simon, 1988).

En general, los vectores que más relación tienen en este estudio se pueden clasificar como mecánicos y biológicos. los primeros del tipo artrópodos hematófagos, son infectados al ingerir la sangre que tiene presente un microorganismo infeccioso, que lo trasmitirá a un nuevo huésped sin pasar por ningún ciclo de aumento del vector, normalmente este tipo de infección que trasmite el vector regularmente son de corta duración. En el segundo tipo de vector los microorganismos infecciosos cuentan con un ciclo de multiplicación dentro del vector que continúan con la infección y además tienen la posibilidad de trasmitir el problema a su descendencia, esta clase de vector biológico

es de especial interés en el campo de la epidemiología dado su gran poder para originar y mantener las enfermedades (S. Zientara, 2015).

Existe una cadena epidemiológica que relaciona el agente infeccioso y el receptor, como toda cadena esta se puede caracterizar como muestra la adaptación de (Sánchez, 2011) en la tabla 4:

Tabla 4. Actores en la Cadena Epidemiológica.

CADENA EPIDEMIOLÓGICA	
TIPO	**DESCRIPCIÓN**
HUÉSPED	Es un organismo que puede albergar parásitos u otros microorganismos.
VECTOR	Es un organismo de tipo animado o inanimado el cual lleva o transmite un parásito u otros microorganismos hacia un huésped.
RESERVORIOS	Es el lugar en que el agente etiológico se mantiene durante un tiempo indefinido.
FUENTES	Es lugar que favorece que el parásito u otros microorganismos se mantengan vivos aun cuando no están en un huésped.

III. ENFERMEDADES TRANSMITIDAS POR VECTORES

Conociendo que los vectores actúan como agentes que transportan algún tipo de virus entre poblaciones de diferentes tipos, el análisis se enfocará hacia los humanos y con los vectores de tipo animado, como por ejemplo algunos que son pertenecientes a la familia flaviviridae y más precisamente el vector distinguido como mosquito, en las siguientes tablas 5, 6 y 7 adaptadas de (Calisher C. H., 2003) se mencionan algunos de los ejemplos de los vectores que componen la familia flaviviridae.

Tabla 5. Vectores del Flavivirus.

FLAVIVIRUS				
VIRUS	**HUÉSPED**	**VECTOR**	**ENFERMEDAD**	**DISTRIBUCIÓN**
Dengue 1-4	Humano	Mosquito	Fiebre, hemorragia	Mundial
Fiebre Amarilla	Primate/Humano	Mosquito	Hemorragia, hepatitis	África, América
Encefalitis Japonesa	Mamíferos, cerdo	Mosquito	Encefalitis	Asia, Australia
Encefalitis de Saint Louis	Mamíferos, aves	Mosquito	Encefalitis	América
Encefalitis de Murray Valley	Mamíferos, aves	Mosquito	Encefalitis	Australia
West Nile	Aves, humanos, mamíferos grandes y pequeños, ganado	Mosquito garrapatas	Fiebre, encefalitis, meningitis, hemorragia, hepatitis.	África, Europa, Asia, América, Australia, Oriente Medio.
Encefalitis transmitida por garrapata	Mamíferos	Garrapata	Encefalitis	Europa, Asia

Tabla 6. Vectores del PestivirusÉ

PESTIVIRUS				
VIRUS	**HUÉSPED**	**VECTOR**	**ENFERMEDAD**	**DISTRIBUCIÓN**
Peste porcina clásica	Cerdo	Contacto	Fiebre, gastroenteritis	Europa, América
Diarrea Bovina	Ganado vacuno	Contacto	Enfermedad	mucosas Mundial

Tabla 7. Vectores del Hepacivirus.

HEPACIVIRUS				
VIRUS	**HUÉSPED**	**VECTOR**	**ENFERMEDAD**	**DISTRIBUCIÓN**
Hepatitis C	Humanos	Parenteral, transfusión	Hepatitis, cáncer de hígado	Mundial

IV. VIRUS TRANSPORTADOS POR VECTORES

Entre las enfermedades que se transmiten por medio de los vectores tipo mosquito se encuentran los que se mencionan en la tabla 8 adaptada según lo publicado por la Organización Mundial De La Salud (OMS, Enfermedades transmitidas por vectores, 2017) .

Tabla 8. Enfermedades transmitidas por mosquitos.

MOSQUITO	
NOMBRE	ENFERMEDADES QUE TRANSMITE
Aedes aegypti	Chikungunya, Dengue, Fiebre del Valle del Rift, Fiebre amarilla, Filariasis linfática, Zika.
Anopheles	Filariasis linfática, Paludismo.
Culex	Encefalitis japonesa, Fiebre del Nilo Occidental, Filariasis linfática.

Complementario con la tabla anterior, estudios recientes de científicos brasileros descubrieron que el virus del zika se transmite no solo por el aedes sino por también por un mosquito común conocido como culex (Duschinka RD Guedes, 2016) lo cual aumenta las alarmas sobre este vector y los virus que puede transmitir, adicionalmente circula un virus llamado usutu que es proveniente de Europa, y a su vez reaparece el virus mayaro que fue identificado en un paciente en Haití donde no existen registros de casos y fue descubierto por investigadores de la Universidad De Florida. Es preocupante que este virus se encuentra rondando el caribe y el norte de Brasil, se transmite por el mosquito aedes lo que generó una gran preocupación de los habitantes del cono sur americano (John Lednicky, 2016). Lo realmente grave es que un solo vector puede infectar a un huésped con varios virus a la vez, es decir si el vector transporta zika, chikungunya y dengue, puede infectar a un solo huésped con los tres virus de manera simultánea según

plantea un estudio científico de la Universidad De Colorado (Waggoner, 2016) y ahora con el potencial resurgimiento del virus mayaro y la amenaza latente del virus usutu, es posible que el receptor pueda ser infectado por todos los virus al mismo tiempo con una sola picadura del mosquito infectado aedes que es capaz de transportar cualquiera de estos virus.

Margaret Chan, directora general de la Organización Mundial de la Salud (OMS), declaró en el año 2016 que el dengue es una amenaza mayor y es más peligroso para los humanos, incluso mayor que el zika, debido a que el dengue puede tener un nivel de complicación mayor. En un reporte publicado en la Organización Panamericana de la Salud presentado por Argentina se estima que el 40% de la población mundial corre riesgo de contraer dengue presentando la posibilidad de llegar a etapas graves de la enfermedad (Dirección de Epidemiología, 2016). Alrededor de medio millón de personas del globo terraqueo contraen dengue al año, de estas el 2.5% mueren a causa de este virus asegura un estudio de la ONU. Con análisis combinados entre multiples fuentes se dice que en la actualidad el dengue es la arbovirosis que promueve el mayor daño en temas de mortalidad (Añez G, 2003), morbilidad (Torres, 2008) y perjuicio económico en zonas afectadas (Donald S. Shepard, 2011).

V. DENGUE

El dengue también conocido como (DEN) es una enfermedad viral que se transmite por la picadura de los mosquitos infectados del tipo hembra que realizan su primera ingesta de sangre 24 horas después de haber emergido y son pertenecientes al género Aedes,

principalmente Aedes aegypti (A. aegypti) y en menor medida el Aedes albopictus que habitan comúnmente en zonas tropicales y subtropicales. Ambos vectores se encuentran distribuidos por toda Colombia, estos tipos de virus son considerados de carácter endemo-epidémico. Las epidemias de dengue documentadas datan desde los años 1779-1780 en Asia, África y América del Norte donde no se conocía la existencia de varios tipos, el dengue cuenta con 4 serotipos reconocidos por la OMS (OMS, Enfermedades transmitidas por vectores, 2017), que son los: DEN-1, DEN-2, que fueron identificados en 1944 y el DEN-3, DEN-4 posteriormente identificados en 1957. Aunque recientemente en octubre de 2013 se encontró una quinta variante de este virus hallada en un paciente en Tailandia y se nombró como DEN-5 (Mustafa, 2015), la forma de clasificar cada tipo de dengue se hace básicamente porque cada virus se diferencia de otro gracias a la secuencia del genoma que lo componen.

La Organización Panamericana De La Salud en su sitio (OPS M. O., 2017) tiene públicos un conjunto de mapas que sirven como apoyo para entender la migración del virus del dengue al continente americano. El primer mapa (la ilustración 1) la forma en que los serotipos 1 al 4 se han ido estableciendo en el continente americano desde la década de los 90 del siglo pasado hasta casi la mitad de la década actual, observando como el sur del continente alberga casi todos los serotipos del virus con detalles de cada serotipo en cada país de la región. El segundo mapa (la ilustración 2) ilustra la incidencia del virus del dengue en el continente desde el año 1980 hasta el año 2014 mostrando las estadísticas de los casos de dengue por país usando la configuración de colores para la interpretación de éste (OMS, Campañas mundiales de salud pública de la OMS, 2017).

Ilustración 1. Distribución serotipos de virus de dengue en las Américas 1990-2014.

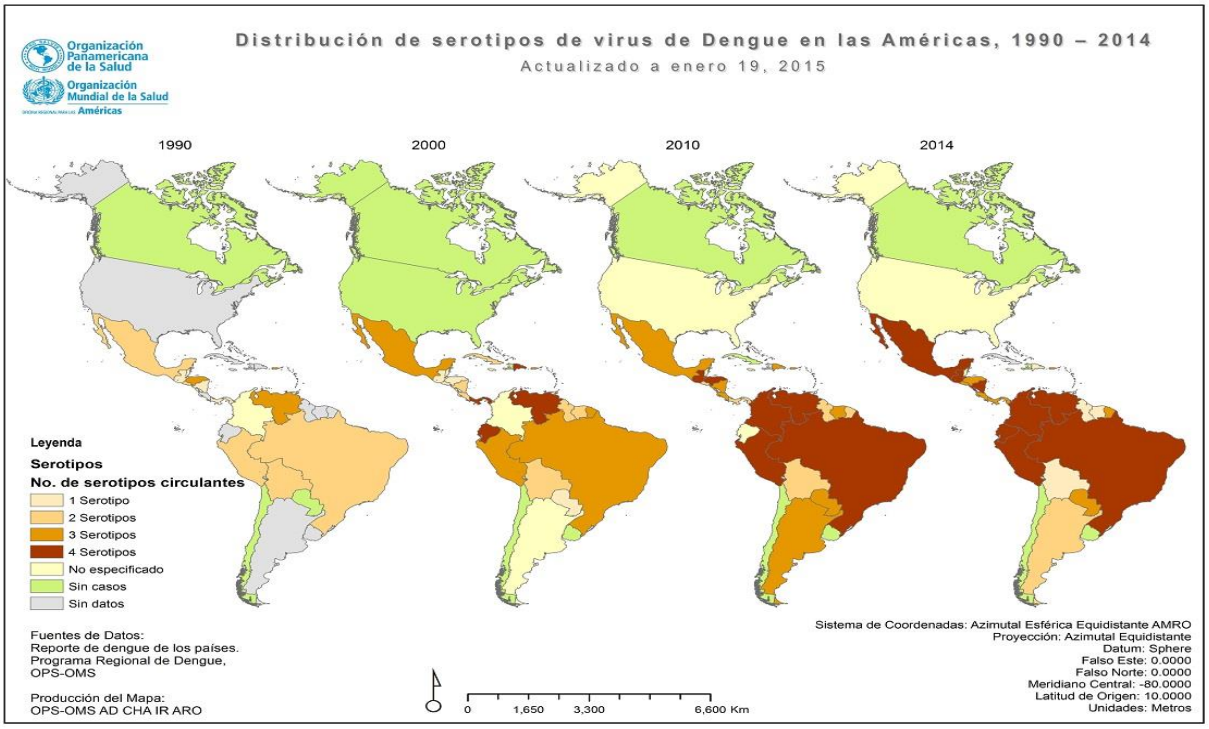
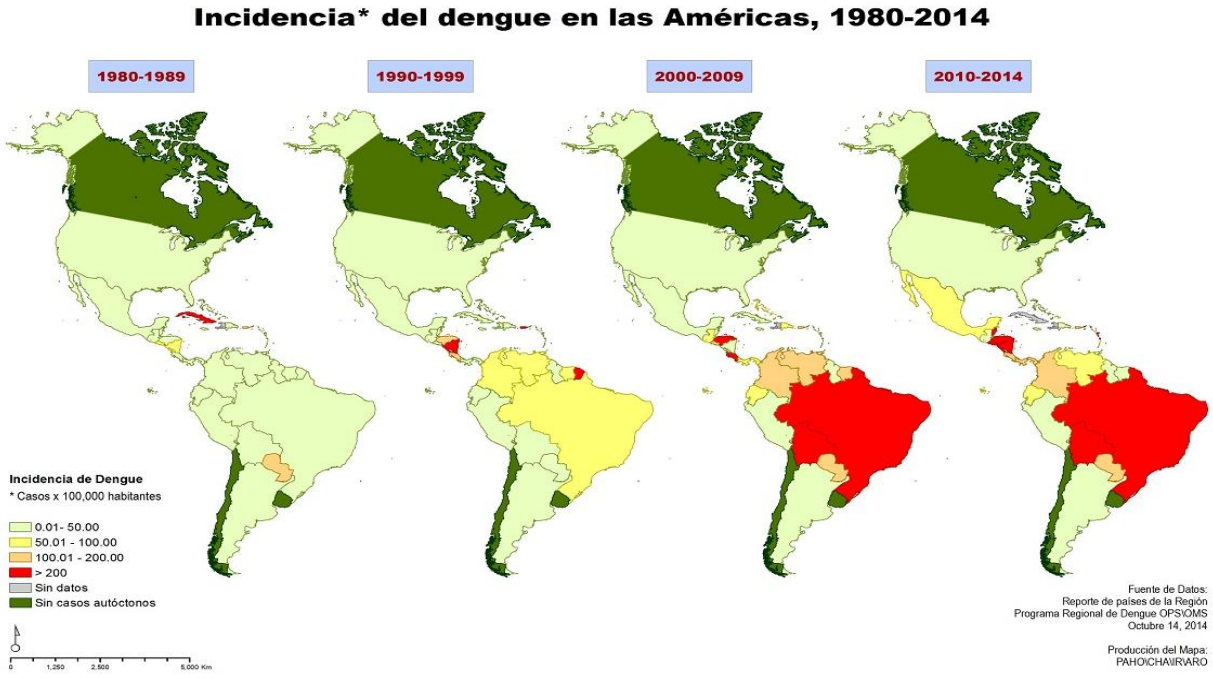


Ilustración 2. Incidencia del dengue en las Américas 1980-2014.



El aedes se convierte en portador e infectante en un periodo contenido entre 8 y 12 días después de ingerir sangre de un sujeto que contenga el virus, desde ese momento el vector se vuelve un gran riesgo debido que permanece infectado por toda su vida e infecta a todo huésped que pique. Los enfermos con el virus, tienen la capacidad de contaminar a los mosquitos vectores en menos tiempo del necesario para finalizar el periodo febril, el cual transcurre regularmente entre tres a cinco días (Quyen NTH1, 2017).

Actualmente se clasifican dos variaciones de este virus, que son el dengue o dengue clásico y el dengue grave que anteriormente se conocía como dengue hemorrágico, estos se manifiestan por tres niveles de fiebre tales como la fiebre de dengue, fiebre hemorrágica de dengue y shock hemorrágico (Urbina, 2017). La Fiebre por dengue se identifica cuando las personas padecen de fiebre y un par de los siguientes síntomas: cefalea, dolor retro-ocular, dolores osteomio-articulares, exantema, leucopenia y sangrado (Nagaram, 2017). La fiebre hemorrágica cuenta con fiebre aguda, sangrado espontáneo: petequias, test positivo del torniquete, trombocitopenia hasta 100.000 células por mm cúbico, sangrado de mucosas, hematemesis, y extravasación de plasma por elevación o disminución del 20% del hematocrito al terminar la etapa crítica, también al comprobar derrame pleural, ascitis o derrame pericardio (Soria Segarra Carmen, 2009). El shock hemorrágico se caracteriza por el constante desgaste del estado del paciente que puede presentar algún carácter de falla circulatoria tales como: taquicardia, cianosis perioral, piel fría con lividez, congestión, acidosis metabólica, convulsiones, hemorragia cerebral y coma (Rodríguez, 2017).

VI. CASOS DE DENGUE

El dengue en Colombia en la década actual se ha mantenido con un promedio de 86.880 casos por año, alcanzando de estos un total de 1.451 en casos de tipo severo y teniendo un índice de muerte de 107 personas anualmente como se muestra en la siguiente tabla adaptada de los reportes que publica cada año la OMS y la OPS (OMS, Enfermedades transmitidas por vectores, 2017). En esta tabla se hace énfasis en los datos reportados por la Organización Panamericana De La Salud en el periodo comprendido entre los años 2008 y la semana 41 del año 2017 que son los que están en el actual reporte, los datos extraídos son los casos de dengue posible y como su contraparte los casos confirmados por año, cada tipo de serotipo del dengue que se encontró a lo largo de esta década, los casos que se convirtieron en dengue severo y por último el total final de las muertes que fueron provocadas por el dengue en cada año de la presente década. Los datos anteriormente mencionados se ven reflejado en las siguientes tablas 9, 10 y 11.

Tabla 9. Casos de dengue en Colombia 2008-2017.

AÑO	CASOS POSIBLES	CASOS CONFIRMADOS	SEROTIPOS DETECTADOS	DENGUE SEVERO	TOTAL MUERTES
2017.41	22.476	6.104	DEN 1,2,3,4	227	44
2016	103.822	45.809	DEN 1,2,3,4	1.047	199
2015	96.444	39.475	DEN 1,2,3,4	1.421	155
2014	105.356	46.842	DEN 1,2,3,4	2.619	166
2013	127.219	65.464	DEN 1,2,3,4	3.377	161
2012	49.361	5.510	DEN 1,3,4	1.329	51
2011	33.207	8.941	DEN 1,2,3,4	1.388	42
2010	157.152	74.763	DEN 1,2,3,4	203	41
2009	51.543	28.503	DEN 1,2,3,4	7.131	44
2008	26.732	3.394	DEN 1,2,3	3.081	12

Un tema de gran preocupación es que Colombia en los últimos 4 años ha reportado el 17.6% de los casos de dengue severo en toda América y en este mismo periodo los casos de letalidad por este virus han alcanzado el 17.2% del reporte continental, como se evidencia en la siguiente tabla adaptada del reporte anual de la OMS y la OPS (Organización Panamericana De La Salud, 2017).

Tabla 10. Comparación Dengue en Colombia contra el continente americano.

AÑO	CASOS DENGUE TODA AMÉRICA	CASOS DENGUE COLOMBIA	% APORTE COLOMBIA A AMÉRICA	DENGUE SEVERO AMÉRICA	DENGUE SEVERO COLOMBIA
2017.4	483.208	6.104	1.3	1.228	227
2016	2.338.848	45.809	2.0	4.274	1.047
2015	2.430.278	39.475	1.6	12.498	1.421
2014	1.176.529	46.842	4.0	16.238	2.619
2013	2.386.836	65.464	2.7	37.898	3.377
2012	1.120.902	5.510	0.5	32.748	1.329
2011	1.093.252	8.941	0.8	19.455	1.388
2010	1.663.276	74.763	4.5	48.954	9.482
2009	1.135.663	28.503	2.5	34.622	7.131
2008	908.926	3.394	0.4	58.521	3.081

Tabla 11. Comparación Muertes por Dengue en Colombia contra el continente americano.

AÑO	DENGUE SEVERO AMÉRICA	DENGUE SEVERO COLOMBIA	%APORTE COLOMBIA A AMÉRICA	TOTAL MUERTES AMÉRICA	TOTAL MUERTES COLOMBIA	% APORTE COLOMBIA A AMÉRICA
2017.4	1.228	227	18.5	253	44	17.4
2016	4.274	1.047	24.5	1.032	199	19.3
2015	12.498	1.421	11.4	1.365	155	11.4
2014	16.238	2.619	16.1	798	166	20.8
2013	37.898	3.377	8.9	1.318	161	12.2
2012	32.748	1.329	4.1	784	51	6.5
2011	19.455	1.388	7.1	763	42	5.5
2010	48.954	9.482	4.5	1.194	41	3.4
2009	34.622	7.131	20.6	618	44	7.1
2008	58.521	3.081	5.3	306	12	3.9

VII. MINERÍA DE DATOS

La minería de datos es un como campo en las ciencias de la computación que se refiere al proceso con el cual se trata de descubrir algún tipo de patrón o patrones cuando se estudian un gran conjunto de datos que contienen un volumen alto de éstos. Este campo usa métodos como la estadística, sistemas de bases de datos, inteligencia artificial y machine learning. La finalidad del proceso de la minería de datos se centra en extraer información de una fuente de datos y obtener una data que sea comprensible para su uso e interpretación posterior. Este proceso se desarrolla en varias etapas consistentes en el análisis crudo de la información, la depuración de los datos, el procesamiento de datos, la creación de un modelo de estudio, y un análisis de deducción de la información, algunas métricas que se busquen evaluar, apreciación basada en la teoría de la complejidad computacional, procesamiento posterior de los descubrimientos, visualización de la información y casi siempre la actualización de los datos en tiempo real (Galit Shmueli, 2017).

El trabajo de la minería de datos se trata en los tipos de análisis semiautomáticos o automáticos de un gran volumen de datos que permiten la identificación de patrones no visibles en la información, que permitan detectar anomalías, dependencias y otra información de interés para el usuario. Regularmente se usan técnicas de base de datos como indexación espacial, al identificar los patrones se pueden utilizar en análisis posteriores, en el machine learning y el análisis predictivo. La minería de datos como todo proceso sigue un conjunto de pasos propuestos para su desarrollo (Aalst, 2016), y son resumidos como: Selección del conjunto de datos, Análisis de las propiedades de

los datos, Transformación del conjunto de datos de entrada, Selección y aplicación de la técnica de minería de datos, Extracción de conocimiento, Interpretación y análisis de datos.

Existen diferentes tipos de minería de datos como son la predictiva y la agrupativa. La primera usa la probabilidad y en algunos casos generan reglas como base para su predicción, la segunda identifica los grupos naturales de los datos y de ahí su agrupación. También la minería de datos cuenta con un par de categorías que son supervisadas y no supervisadas (Pei, 2011).

En el aprendizaje supervisado es necesario tener un conocimiento previo que permita tener una referencia y por eso también se le llama aprendizaje dirigido debido a que se basa en un atributo conocido y este sirve como modelo base, es por eso que el aprendizaje dirigido regularmente genera modelos predictivos. Contrario a esto el aprendizaje no supervisado no es dirigido, por eso no existe distinción entre los atributos dependientes y no dependientes, debido a esto no se puede prever el resultado que ayude a que un algoritmo se favorezca en la construcción de un modelo, la aplicación más usada es para temas descriptivos, pero en algunas ocasiones se podría utilizar para hacer algún tipo de predicción (Okun, 2010).

VIII. MACHINE LEARNING

El aprendizaje sobre la información ha sido usado desde hace mucho tiempo, como por ejemplo los filósofos la llaman inferencia inductiva, ni siquiera en el siglo XX se consideró que fuera posible la inducción pura sin que antes se tenga conocimiento del tema específico. Este caso ha sido estudiado por mucho tiempo, Gauss en el siglo XVIII propuso desde la perspectiva estadística la idea de usar la regresión de mínimos cuadrados, ya desde la década de los años 30, del siglo anterior, con la aproximación para la clasificación de Fisher se mantiene como la base para el desarrollo de análisis y métodos (Shawe-Taylor, 2007).

La idea de máquinas de aprendizaje fue propuesta en 1950 por Alan Turing con el propósito de modelar los problemas de aprendizaje como problemas de tipo búsqueda en una zona de hipótesis adecuada, principalmente se buscaba como enfoque para llevar a cabo la inteligencia artificial. Por esta razón la creación de algoritmos de aprendizaje se convirtió en un campo vital para la inteligencia artificial, posteriormente se hizo tan importante el campo del aprendizaje que se creó su área propia conocida como machine learning, eventualmente formando el área separada de aprendizaje de máquina.

Existen varios tipos de aprendizaje de máquina: aprendizaje supervisado, no supervisado, aprendizaje reforzado, aprendizaje semi-supervisado, aprendizaje inductivo, aprendizaje activo, entre otros. Un sistema para ser considerado inteligente debe tener la capacidad de aprender de manera automática ya que esto es fundamental en la inteligencia artificial y éste es el principio del funcionamiento del machine learning

(Miroslav Kubat, 1988), su finalidad es la creación de métodos computacionales que permitan la implementación de múltiples formas de aprendizaje. Resumiendo, la tarea principal del machine learning se puede explicar cómo el proceso automático en el cual se trata de obtener una mejor solución que se les da a los problemas. Los métodos de que se usan en machine learning se pueden agrupar de forma muy general como inductivos y no inductivos (Carbonell, 1990).

En el aprendizaje de máquina o machine learning se tiene el aprendizaje supervisado, el aprendizaje no supervisado, y el aprendizaje por refuerzo, los primeros dos tipos de aprendizaje funcionan similar a los empleados en la minería de datos, y el último tipo aprendizaje por refuerzo usa el aprendizaje por recompensas basadas en la ley del efecto, donde las repuestas que son seguidas de consecuencias reforzantes se asociarán al estímulo y tendrán mayor probabilidad de ocurrencia cuando el estímulo reaparezca, entonces las acciones que tienen como resultado una consecuencia o meta positiva son aprendidas (Thorndike, 1927) y aunque fue una teoría planteada por un psicólogo en la década de los años 20 del siglo pasado, esta planto las bases del aprendizaje por refuerzo en el machine learning, donde la maquina aprende los caminos que generan mejores respuesta e ignora los caminos que dan malas respuestas (Richard Sutton, 1998). También es de considerar que existe el aprendizaje semi-supervisado, el aprendizaje inductivo el aprendizaje activo entre otros (Alpaydin, 2010).

3. ESTADO DEL ARTE

I. EVOLUCIÓN COLOMBIANA DE LOS VECTORES

Se presume que la forma en que el mosquito Aedes llegó a Colombia fue por intermedio de los barcos de esclavos que venían provenientes de África y que embarcaban en Cartagena y la manera como comenzó a dispersarse hacia el interior del país fue gracias a la navegación sobre el río Magdalena, se tornó importante debido a que la fiebre amarilla urbana estaba presente en los puertos más importantes del continente americano. Para el año de 1880 el mosquito fue identificado en la ciudad de Neiva, se presume que en 1883 el Aedes aegypti llegó a Cúcuta proveniente de Maracaibo Venezuela, quien sufría una epidemia de fiebre amarilla, el mosquito se convirtió en habitante del norte de Santander y causó las epidemias hasta 1912. En 1910 y 1923 se generaron epidemias de fiebre amarilla urbana en la ciudad de Bucaramanga, en 1929 en Socorro Santander, hubo un brote epidémico fuerte de fiebre amarilla urbana, esas epidemias de fiebre amarilla se lograron controlar de gran manera ya que se actuó frente al vector que lo producía. En 1915 llegó el vector a puerto Berrio y Cisneros en Antioquia; en 1923 se realizó una encuesta en Cúcuta que permitió entender la magnitud de infestación del vector, que consiguió hasta un 90% de presencia en las casas de la ciudad, ya en el año 1926 si hicieron grandes campañas para erradicar el vector, con una increíble disminución de este, 90% al 0.1% con la ayuda del petróleo y peces larvivoros que se usaban en tanques elevados (Galvis, 1982).

Por la mitad de la década de 1950 el dengue era de tipo endémico ya que tenía infectado el 28% del territorio colombiano (Boshell J, 1986), en la costa atlántica, los valles

interandinos de los ríos Magdalena y Cauca, en Buenaventura y Cúcuta, se calculaba que un área aproximadamente de 352.507 Km^2 estaban bajo la influencia del vector y tenía una población con potencial de infección de 7.193.310 personas (Groot, 1980). El país se liberó del mosquito y del dengue por dos décadas, pero en 1960 en Cúcuta se descubrió una cepa resistente a los controles y pese a los esfuerzos por erradicarlo, éste no pudo ser eliminado (Morales, 1991). En 1968 hubo una reaparición del dengue en Maracaibo que afectó a la Guajira y con esto se generó una reinfección del dengue en toda la costa Atlántica, para 1971 y 1972 se cree que por la interacción de pobladores infectados de la frontera entre Colombia y Venezuela se produjo una epidemia que sumó a 500.000 personas (OPS, Guía de informes de la Campaña de Erradicación del Aedes aegypti en las Américas, 1960), se reinsertó el mosquito en el interior del país por la misma vía de la vez anterior utilizando el rio Magdalena, reconquistando Santander y el departamento del norte de Santander para el año 1975, en el año 1975 se expandió a Caquetá, Meta, Tolima y Cundinamarca, para el año 1981 alcanzó el Vaupés, en el año 1997 a excepción del Vaupés, Caquetá y Guainía, el Aedes aegypti estaba disperso en toda la nación (H Groot, 1976).

El mosquito aedes una vez infectado continua así por el resto de su vida que es de un promedio de 30 días, también el género que más ataca del vector al hombre es el tipo hembra, esta hace su primera ingesta de sangre con solo 24 horas después de haber emergido, el mosquito cuando es adulto tiene unas pintas en el dorso color plata o amarillento, el ultimo arteto de las patas tiene pintas de color blanco y en las hembras el abdomen tiende a tomar una forma puntiaguda, la principal especie aedes aegipty

evoluciona en dos diferentes etapas en su vida, la primera parte es la fase acuática donde se desarrolla como huevo, larva y pupa, en esta fase los huevos han logrado resistir condiciones como la desecación y en la segunda fase es la fase de adulto o fase aérea (Fernández, 1999). El mosquito mantiene regularmente la actividad de alimentación justo cuando la luz solar cuenta con poca intensidad como en el comienzo del amanecer o antes de que se extinga la luz del atardecer. Pero también este vector puede atacar en el día o en la noche dependiendo de la cantidad de sangre que necesite. Tiene un ciclo para poner huevos de alrededor de 3 días logrando producir hasta 200 huevos, los huevos completan su desarrollo de tipo embrionario en solo 48 horas si cuentan con las condiciones óptimas que son temperatura entre 24 y 28 grados centígrados y hasta un 70% de humedad relativa, o pueden tardar hasta cinco días si las temperaturas son bajas (JS Zuluaga, 2002). Este vector holometábolo prefiere las zonas urbanas, estar cerca de la población humana, especialmente en habitaciones o sitios que de preferencia los protejan y oculten de la luz como armarios, baños, bajo los enceres y en jardines, materos, etc.

El ciclo completo del vector es de la siguiente manera: los huevos se encuentran en recipientes o lugares que tengan agua estancada, se sabe que la fase del huevo se tarda hasta dos días en completarse, posteriormente se llega a la fase de larva en la que se alimenta de todo material orgánico que se encuentre en los recipientes donde se encuentren, esta fase dura entre cinco y siete días si cuentan con temperaturas entre 13 y 45 grados Celsius, luego de ésto pasan al estado de pupa que dura entre uno y dos días, en esta fase no se alimentan y sufren cambios fisiológicos y anatómicos hasta que

se convierten en adulto y de ahí abandonan el agua para salir a las zonas donde estén sus fuentes de alimento como los humanos (Reyes-Villanueva, 1990). Los efectos de la infección del dengue en los humanos presentados por los diferentes serotipos pueden ser similares, cuando una persona es infectada por el virus del dengue al recuperarse el organismo se vuelve inmune al virus, pero la inmunidad que se obtiene al ser infectado por un serotipo en particular no es protectora por parte de los demás serotipos, es decir si una persona es infectada con el virus DEN-1 será inmune a este serotipo una vez se recupere, pero si un mosquito tiene el virus DEN-2 e infecta a una persona que no haya sufrido bajo la influencia de ese serotipo especifico, ésta será infectada por el virus y al recuperarse ya tendrá los anticuerpos que le permitan evitar un nuevo contagio del DEN-2. Es de tener en cuenta que el ser humano es el único organismo conocido que es capaz de hospedar el virus del dengue y ser infectado por éste (L Morier, 2000).

El virus del dengue proviene de la familia flaviviridae de los arbovirus, son virus esféricos con una capa protectora de 40-50 nanometros (nm) como diámetro con cápside icosaédrica y genoma de ácido ribonucleico monocatenario, es no segmentado y cuenta con polaridad positiva. Este genoma trabaja directamente como el ácido ribonucleico monocatenario mensajero policistronico, el virión se compone de una nucleocápside de simetría cúbica, rodeada bajo una estructura lipoproteíca que proviene de la célula que es hospedera, Tener este tipo de estructura lo convierte en lábil a la desactivación por solventes de tipo orgánicos, a los cambios de pH y cambios de temperatura (Rothman, 2004). El genoma se compone de una única molécula de ácido ribonucleico con cadena lineal sencilla compuesto de 10.703 nucleótidos y gran variación genómica. Los ácidos

nucleicos genómicos del virus son infecciosos por sí mismos, dado esto se hace necesario tratar este virus con el segundo nivel de bioseguridad tal como lo recomiendan las autoridades de salud, el virus cuenta con la capacidad de adherirse a las células, ensamblarse en retículo endoplasmático y multiplicarse en el citoplasma, el genoma de este codifica una poliproteina que posteriormente se procesa en 10 polipéptidos, 3 de tipo estructural y 7 de tipo no estructural (Halstead, 2002).

Los cuadros clínicos del dengue son diversos, pero para ayudar con un diagnóstico más ágil la OMS y la OPS, crearon una clasificación de este virus en cuatro grados de intensidad, los dos primeros grados corresponden a la fiebre del dengue y el dengue hemorrágico, los grados 3 y 4 son los que corresponden al choque por dengue. El grado 1 se identifica con fiebre que puede tener adicionalmente síntomas no específicos, la forma de identificarlo es usando la prueba del torniquete positiva. El grado 2 tiene como síntomas los del grado uno, sumándole hemorragias espontáneas, casi siempre hemorragia cutánea, de otros lados del cuerpo o los dos al tiempo. El grado 3 los síntomas son pulso rápido y débil, hipotensión, agitación, piel fría y húmeda. El grado 4 pulso tan bajo que es imperceptible, choque profundo con la presión arterial (J.G RIGAU PEREZ, 1999).

Los casos de dengue hemorrágico, dengue por choque o dengue por choque profundo se presentan con regularidad cuando la persona tiene en su cuerpo dos o más serotipos, sea porque los adquirió al mismo tiempo o porque tuvo un serotipo anteriormente y ahora fue infectado por otro serotipo que no había sufrido, el dengue en la actualidad no cuenta

con una vacuna, los tratamientos contra este virus por deshidratación, debido al vómito, la fiebre alta y la poca ingesta de alimentos son consumir un gran volumen de líquidos mediante vía oral, la OMS recomienda una solución que permite rehidratar al infectado que tiene síntomas diarreicos fuertes que está compuesta por 3.5 gramos de cloruro de sodio, 1.5 gramos de cloruro de potasio, 2.5 gramos de carbonato acido de sodio y 20 gramos de glucosa todos disueltos en un litro de agua (OMS W. H., 1997). Si en la fase febril existe posibilidad de convulsiones por hiperpirexia, es posible administrar un antipirético como el acetaminofén solo si la temperatura corporal es superior a los 39 grados Celsius, sin superar las seis dosis en un intervalo de 24 horas. Se recomienda no usar silicatos porque estos bloquean la agregación de las plaquetas e interfieren en la formación de protombina lo que puede complicar al paciente si este presenta un cuadro hemorrágico.

Cuando el paciente es un caso de dengue hemorrágico o dengue por choque es necesario actuar rápido porque el suplir la perdida de líquidos es vital para la supervivencia del infectado, por esto con la administración de electrolitos, soluciones, plasma y sangre se puede detener el cuadro de dengue por choque rápidamente y con esto se evita la coagulación intravascular diseminada. El pronóstico sobre un paciente es dependiente de lo rápido que se identifique el dengue precoz, eso hace necesario un especial seguimiento y control de estos casos donde el periodo crítico ocurre cuando se pasa de la fase febril a la afebril que en lo general sucede a partir del tercer día de iniciar el cuadro clínico. Para tratar esos cuadros de dengue hemorrágico y dengue por choque se realiza una solución salina con glucosa al 5% y ringer de lactato la cual se administra

por vía intravenosa al paciente, si presenta caso de acidosis metabólica se prepara una solución básica con bicarbonato de sodio y una concentración de 0.167 moles por litro de solución. Si el dengue de choque se agrava a choque profundo se administra plasma o el dextrano 40 que actúa como sustituto del plasma, para cualquiera de los casos de administración de líquidos intravenosos se debe administrar por un máximo de 48 horas posterior a la recuperación del valor del hematocrito y los signos vitales, es necesario tener un control precavido para evitar la aparición de un edema pulmonar, insuficiencia cardiaaca o hipervolemia (OMS W. H., 1997).

II. ALGORITMOS DE MACHINE LEARNING

Con los brotes de las epidemias que han surgido en la última década, especialmente en el cono sur de América, donde Colombia no está exenta y es de los que más se ve afectada por este flagelo, es importante buscar un método que permita encontrar algunos indicios sobre los factores que pueden generar un problema de salud pública local, regional o nacional, entendiendo los avances tecnológicos que existen en la actualidad sobre todo en el campo de la computación es posible aprovechar estos recursos para evaluar las situaciones como las de este tipo de problemas y aprender de éstas usando la inteligencia artificial como aliada y específicamente técnicas de aprendizaje automático o machine learning que permitan encontrar algún tipo de coincidencias en los problemas que se presentan usando algoritmos sobre los datos del estudio, logrando de alguna manera identificar los causantes y poder tomar medidas para actuar sobre los focos que pueden estar ocasionando el problema.

Antes de mencionar algunos de los algoritmos que se usan para la minería de datos y el machine learning es necesario tener claro la definición de algoritmo que el autor Luis Joyanes Aguilar propone como “una secuencia ordenada de pasos, sin ambigüedades, que conducen a la solución de un problema dado” (Joyanes Aguilar Luis, 1996), analizando esta definición se podría decir que los algoritmos se construyen con la finalidad de conseguir un objetivo que a su vez se encuentra relacionado con el cómo lograrlo y los resultados en forma cuantificable, cualificable o ambas que se esperan obtener al realizar la aplicación de éste. Por esto, se hace necesario construir un algoritmo propio si es el caso o elegir un algoritmo preciso, finito y completamente definido. Como se había mencionado anteriormente el machine learning cuenta básicamente con tres tipos que son el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo. En la actualidad no se ha logrado conseguir que una computadora aprenda tan bien como lo hace un humano, algunos algoritmos que existen en la actualidad y que pertenecen a este campo han conseguido avances interesantes en algunas tareas del campo del aprendizaje, como por ejemplo los algoritmos de clasificación.

En este tipo de algoritmos se busca hacer la aproximación de una función objetivo que no se conoce $\Phi : \mathrm{I} \times \mathrm{C} \rightarrow \{\mathrm{T}, \mathrm{F}\}$ (esta describe instancias de un problema que deben ser evaluadas y clasificadas por una persona experta en ese dominio) a través de otra función $\Theta : \mathrm{I} \times \mathrm{C} \rightarrow \{\mathrm{T}, \mathrm{F}\}$ que se conoce como clasificador en donde $C = \{c1, K, c\ c\}$ pertenece a un conjunto de categorías establecidas, y la I pertenece a un conjunto de instancias del problema. Normalmente cada instancia $i_j \in I$ se representa como una lista

$A = \{a_1, a_2 \ldots, a_{|A|}\}$ de valores que son conocidos como los atributos $i_j = \{a_{1j}, a_{2j} \ldots, a_{|A|j}\}$. Cuando se cumple que $\Phi : i_j \times c_i \rightarrow T$ se le conoce a esto como ejemplo positivo de c_i pero si se cumple que $\Phi : i_j \times c_i \rightarrow F$ se le conoce como ejemplo negativo de c_i.

Cuando se busca generar un clasificador automático de c_i se hace necesario un proceso inductivo conocido como el aprendiz, este obtiene los atributos que una instancia no conocida debe cumplir para hacer parte de la categoría lo cual se logra al observar los atributos de una serie de instancias preclasificadas bajo c_i o no c_i. Por esta, razón al construir el clasificador se hace necesario tener disponible una colección Ω inicial de ejemplos para los cuales el valor $\Phi(i_j, c_i)$ se conozca para todo $(i_j, c_i) \in \Omega \times C$ a un tipo de colección se le conoce como conjunto de entrenamiento (Tr), como conclusión a este proceso se la llama aprendizaje supervisado dado que depende de (Tr). La calidad de un clasificador se mide con la efectividad que este brinde, dado que con esto se logra evaluar su confiabilidad al comparar metodologías, y aunque la eficiencia es importante esta queda relegada por la efectividad debido a que la eficiencia tiene una fuerte dependencia con parámetros, hardware y software. La exactitud y el porcentaje de aciertos son los mejores atributos para evaluar un clasificador. Una manera de calcular la exactitud es en términos de la tabla de contingencia así:

$$\alpha_i = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i}$$

Con el cual se dice que $\mathrm{FP_i}$ se definen como los falsos positivos y corresponde al número de instancias que fueron incorrectamente clasificados bajo c_i, $\mathrm{TP_i}$ son verdaderos positivos, $\mathrm{TN_i}$ son verdaderos negativos y $\mathrm{FN_i}$ son falsos negativos.

Cuando se va a realizar un análisis es normal partir a Ω en tres conjuntos diferentes, un conjunto de entrenamiento, un conjunto de prueba y un conjunto de validación. El conjunto de entrenamiento es con el que se pretende que el aprendiz fabrique su clasificador basado en un conjunto de instancias observadas, el conjunto de validación es aquel con el que se busca optimizar los parámetros en los clasificadores o con el fin de elegir uno en particular, y el conjunto de pruebas es aquel que sirve para definir qué tan efectivo es el clasificador que se fabricó para el análisis. Esta técnica tiene una limitación cuando el volumen de datos en Ω es limitado, para solventar esto y calcular correctamente la exactitud de la técnica de machine learning se debe usar la validación cruzada de 10 pliegues (I. Witten, 2000), aquí se plantea que es necesario partir Ω en 10 conjuntos de manera aleatoria pero manteniendo cada parte la proporción original de cada clase, seguidamente una parte se mantiene fuera por un instante y la técnica de aprendizaje entrena con las nueve partes adicionales, así la exactitud se calcula sobre la parte que no participó en el entrenamiento, y el ciclo sigue hasta que cada parte deje de participar en el proceso de entrenamiento, de ahí la exactitud calculada en los diez procesos de entrenamiento es promediada para tener una estimación completa de la misma, también existen otras maneras de medir la exactitud de un clasificador las cuales son dependientes de algún tipo de aplicación específica.

III. INVESTIGACIONES DE REFERENCIA

Para la realización de este trabajo se tendrán en consideración varios factores al momento de la elección de los algoritmos que se usarán finalmente para el estudio de la información, por esto se citan varios trabajos entre investigaciones de tipo científica, académica y tecnológica, así como tesis doctorales en los que todos los trabajos expuestos tienen relación para alcanzar la meta de esta tesis y que se mencionan como de tipo fichas bibliográficas, identificando su autor, palabras clave, fuente, resumen original del artículo, libro o tesis y los aspectos que se consideran que sirven como aporte para esta tesis y el autor las nombra a continuación.

TÍTULO: Developing a dengue forecast model using machine learning: A case study in China.

PALABRAS CLAVE: machine learning, dengue, forecast, tropical Diseases, China.

FUENTE: Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., ... & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. PLOS Neglected Tropical Diseases, 11(10), e0005973.

RESUMEN: In China, dengue remains an important public health issue with expanded areas and increased incidence recently. Accurate and timely forecasts of dengue incidence in China are still lacking. We aimed to use the state-of-the-art machine learning algorithms to develop an accurate predictive model of dengue.

APORTE A LA TESIS: La fundamentación teórica que brinda esta investigación al usar un gran conjunto de técnicas para el tratamiento y valuación de los datos, y el uso de algoritmos de machine learning que se adecuaban muy consistentemente al análisis propuesto, permiten dar noción de la manera de cómo podría enfrentarse un análisis que contenga las variables de estudio en esta tesis, ya que haciendo una mezcla de verificación de los datos obtenidos posterior a un análisis con otro algoritmo puede mejorar el entendimiento de la información que es objeto de estudio.

TÍTULO: Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore.

PALABRAS CLAVE: machine learning, dengue, forecast, models, decision support, demography.

FUENTE: Shi, Y., Liu, X., Kok, S. Y., Rajarethinam, J., Liang, S., Yap, G., ... & Lo, A. (2016). Three-month real-time dengue forecast models: an early warning system for outbreak alerts and policy decision support in Singapore. Environmental health perspectives, 124(9), 1369.

RESUMEN: With its tropical rainforest climate, rapid urbanization, and changing demography and ecology, Singapore experiences endemic dengue; the last large outbreak in 2013 culminated in 22,170 cases. In the absence of a vaccine on the market, vector control is the key approach for prevention.

APORTE A LA TESIS: Los modelos para pronóstico de dengue que se proponen al tener como referencia el vector Aedes aegypti, para un periodo definido en esta investigación con la finalidad de prevenir de manera oportuna los brotes y la posibilidad de epidemias en Singapur, con la información en cuanto que la demografía, ecología y clima tropical similar a Colombia, es interesante tener una buena referencia con condiciones similares a la del país en estudio para tomar en cuenta al momento de la evaluación de los datos de esta tesis, además los algoritmos que usan para el estudio de los datos como LASSO refuerzan de alguna manera la elección de cuáles de estos deben ser considerados para el estudio de los datos.

TÍTULO: Forecast of dengue incidence using temperature and rainfall.

PALABRAS CLAVE: time series, Poisson multivariate regression model, dengue.

.

FUENTE: Hii, Y. L., Zhu, H., Ng, N., Ng, L. C., & Rocklöv, J. (2012). Forecast of dengue incidence using temperature and rainfall. PLoS neglected tropical diseases, 6(11), e1908.

RESUMEN: An accurate early warning system to predict impending epidemics enhances the effectiveness of preventive measures against dengue fever. The aim of this study was to develop and validate a forecasting model that could predict dengue cases and provide timely early warning in Singapore.

APORTE A LA TESIS: Se abordó como teoría que factores climáticos como la lluvia y la temperatura están relacionados de alguna manera con el virus del dengue, entonces proponen que las epidemias del virus del dengue pueden ser anticipadas cuando se conoce el clima, que también corresponde a un tipo de predicción basada en probabilidades que aún no alcanza el 100% de confiabilidad, usar técnicas como la auto regresión, la estacionalidad y la tendencia para el modelo, entonces tener en consideración que las variables climáticas son importantes y usar las técnicas para identificar y validar el estudio es un factor que pueda ser determinante para las conclusiones de la tesis.

TÍTULO: The ocurrence of dengue and weather changes in Brazil: a systematic review.

PALABRAS CLAVE: machine learning, dengue, forecast, models, decision support, demography.

FUENTE: Viana, D. V., & Ignotti, E. (2013). The ocurrence of dengue and weather changes in Brazil: a systematic review. Revista Brasileira de Epidemiologia, 16(2), 240-256.

RESUMEN: Dengue is configures in recent decades as an important cause of morbidity and mortality in Brazil and around the world reaching the tropical and subtropical areas. To review the scientific literature on the occurrence of dengue in Brazil and its relationship with meteorological variables.

APORTE A LA TESIS: Una revisión profunda del estado del arte en publicaciones de nivel mundial donde se propone que variables como el clima, la humedad, la presión y hasta la dirección de los vientos en una región como la de Brasil en el estudio, tienden a favorecer la expansión del virus del dengue. Dado que Colombia cuenta con regiones tan diversas en clima, presión atmosférica y altura sobre el nivel del mar, pueden intensificar la multiplicación del vector que transmite la enfermedad, el artículo menciona que por el cambio climático para el año 2085 se tiene estimado que el 60% de la humanidad entre 5 a 6 mil millones de personas estarán en riesgo de contraer el virus del dengue, tomar como referencia los estudios que se abordan en este artículo adicionan puntos de vista para tener en cuenta en las variables de análisis de los datos de la tesis.

TÍTULO: Evaluating probabilistic dengue risk forecasts from a prototype early warning system for Brazil.

PALABRAS CLAVE: brazil**,** dengue, forecast, models, early warning, world cup, prototype.

FUENTE: Lowe, R., Coelho, C. A., Barcellos, C., Carvalho, M. S., Catao, R. D. C., Coelho, G. E., ... & Rodo, X. (2016). Evaluating probabilistic dengue risk forecasts from a prototype early warning system for Brazil. *Elife*, *5*, e11285.

RESUMEN: Recently, a prototype dengue early warning system was developed to produce probabilistic forecasts of dengue risk three months ahead of the 2014 World Cup in Brazil. Here, we evaluate the categorical dengue forecasts across all microregions in Brazil, using dengue cases reported in June 2014 to validate the model. We also compare the forecast model framework to a null model, based on seasonal averages of previously observed dengue incidence. When considering the ability of the two models to predict high dengue risk across Brazil, the forecast model produced more hits and fewer missed events than the null model, with a hit rate of 57% for the forecast model compared to 33% for the null model. This early warning model framework may be useful to public health services, not only ahead of mass gatherings, but also before the peak dengue season each year, to control potentially explosive dengue epidemics.

APORTE A LA TESIS: En un evento tan importante a nivel mundial como La Copa Mundo de Futbol donde la migración de personas hacia Brasil es bastante alta por los aficionados a este deporte, aumenta el riesgo de multiplicar los casos de dengue en ese país y también el de la migracion de los serotipos de virus a otros países y regiones, fabricar un modelo de pronóstico con otro modelo llamado nulo y compararlos entre ambos para definir cuál de los dos muestra mejores resultados, ayuda a tener una referencia otra técnica para la evaluación de los datos para obtener un modelo de probabilidad aceptable que se acople a las necesidades del estudio después de un análisis y no desecharlo si los resultados no son cercanos al 100% de éxito.

TÍTULO: Prediction and prevention of dengue epidemics.

PALABRAS CLAVE: education, dengue, virus, Breteau and/or positive house index, number of indoor resting Aedes females, environmental factors.

FUENTE: Fauran, P. (1996). Prediction and prevention of dengue epidemics. Bulletin de la Societe de pathologie exotique (1990), 89(2), 123-6.

RESUMEN: Prediction and prevention of dengue epidemics are based on informations gathered about the mosquito vector species, the dengue types transmitted, the vertebrate hosts and their environment. Although Aedes aegypti is the most important vector, other Aedes may also propagate the dengue viruses. The populations of vector mosquitoes are evaluated through several indices: Breteau and/or positive house index, number of indoor resting Aedes females, etc.... The four dengue types can replicate in vertebrate hosts beside humans and in other mosquito species than Ae. aegypti. The incidence of dengue on a population is largely variable according to the immunity status, the vector competence and the virus strains. Concomitant infections by two types of dengue virus or by an another pathogen (Alphavirus) have been observed. The environmental factors influencing the dengue ecosystem are mostly climatic (temperature, rainfall, wind) but also anthropic (transportation means, public buildings). Prevention of dengue epidemics must be based on public health education in schools, community participation, epidemiological surveillance linked with good vector control teams. Nevertheless

intensive research on dengue and the actions undertaken for the last forty years, dengue remains the first cause of viral morbidity worldwide.

APORTE A LA TESIS: El factor de la educación es una variable adicional que se suma a otras como las de tipo climático, no solo basta analizar unos datos con algoritmos que permitan evaluar y obtener patrones característicos de tipo para poder predecir una epidemia, sino se tiene en cuenta variables adicionales como por ejemplo la educación de las personas en temas de salud pública. La falta de este ítem en los centros educativos también hace parte de que el vector se recrudezca en las zonas propensas a infecciones dada la poca capacitación que se rinda para las campañas anti vectores que permitan la prevención y eliminación de este flagelo, también es necesario considerar evaluar con índices como breteau y/o índice positivo de la casa, la población de los mosquitos vectores que están en una zona dada en especial el número de hembras Aedes dado que la incidencia del dengue en una población es en gran medida variable de acuerdo con el estado de inmunidad, la competencia del vector y las cepas del virus.

TÍTULO: Mapping the spatial distribution of Aedes aegypti and Aedes albopictus.

PALABRAS CLAVE: Global distribution, Aedes aegypti, Aedes albopictus, Multidisciplinary datasets, Machine learning models.

FUENTE: Ding, F., Fu, J., Jiang, D., Hao, M., & Lin, G. (2017). Mapping the spatial distribution of Aedes aegypti and Aedes albopictus. Acta Tropica.

RESUMEN: Mosquito-borne infectious diseases, such as Rift Valley fever, Dengue, Chikungunya and Zika, have caused mass human death with the transnational expansion fueled by economic globalization. Simulating the distribution of the disease vectors is of great importance in formulating public health planning and disease control strategies. In the present study, we simulated the global distribution of Aedes aegypti and Aedes albopictus at a 5×5km spatial resolution with high-dimensional multidisciplinary datasets and machine learning methods Three relatively popular and robust machine learning models, including support vector machine (SVM), gradient boosting machine (GBM) and random forest (RF), were used. During the fine-tuning process based on training datasets of A. aegypti and A. albopictus, RF models achieved the highest performance with an area under the curve (AUC) of 0.973 and 0.974, respectively, followed by GBM (AUC of 0.971 and 0.972, respectively) and SVM (AUC of 0.963 and 0.964, respectively) models. The simulation difference between RF and GBM models was not statistically significant ($p>0.05$) based on the validation datasets, whereas statistically significant differences ($p<0.05$) were observed for RF and GBM simulations compared with SVM simulations. From the simulated maps derived from RF models, we observed that the distribution of A. albopictus was wider than that of A. aegypti along a latitudinal gradient. The discriminatory power of each factor in simulating the global distribution of the two species was also analyzed. Our results provided fundamental information for further study on disease transmission simulation and risk assessment.

APORTE A LA TESIS: El uso de herramientas como los sistemas de información geográfica para evaluar la distribución del Aedes aegypti en un contexto local, regional o global, para fabricar modelos biológicos basados en la temperatura y poder construir mapas de idoneidad para la subsistencia de los vectores, apoyado en una forma alterna como la distribución probabilística basada en el área bajo la curva, que además en este estudio tuvieron en cuenta una nueva variable importante que fue la cubierta del suelo de las zonas afectadas, también la transformación de los datos que tenían en un sistema de coordenadas único para el análisis, aprovecharon además una variable poco común en los análisis de datos como la ausencia de casos en algunas zonas y la incorporación del algoritmo de Random Forest para la comparación de resultados contra los algoritmos SVM y GBM, entender que detalles tan sutiles como el verdor del dosel de la vegetación es importante para el vector, toda esa serie de conceptos y métodos para evaluar los factores del vector pueden ser aplicados al territorio colombiano dado la diversidad con que se cuenta y que en esos pequeños detalles se puede encontrar la clave para la prevención y por qué no la erradicación definitiva del vector.

TÍTULO: State-of-the-art monitoring in treatment of dengue shock syndrome: a case series.

PALABRAS CLAVE: Machine learning, Decision support, Pulse oximetry, Dengue shock síndrome.

FUENTE: Moulton, S. L., Mulligan, J., Srikiatkhachorn, A., Kalayanarooj, S., Grudic, G. Z., Green, S., ... & Thomas, S. J. (2016). State-of-the-art monitoring in treatment of dengue shock syndrome: a case series. Journal of medical case reports, 10(1), 233.

RESUMEN: Early recognition and treatment of circulatory volume loss is essential in the clinical management of dengue viral infection. We hypothesized that a novel computational algorithm, originally developed for noninvasive monitoring of blood loss in combat casualties, could: (1) indicate the central volume status of children with dengue during the early stages of "shock"; and (2) track fluid resuscitation status.

APORTE A LA TESIS: El análisis con el uso de el algoritmo Compensatory Reserve Index, en el diagnóstico y posterior de control de una versión más fuerte y peligrosa del virus como es el dengue hemorrágico por choque que permite evaluar de manera no invasiva a las personas que integran diferentes tipos de variables tanto fisiológicas, como las del cambio de estado de las personas y características clínicas que permiten predecir de manera personalizada por afectado. Este tipo de algoritmo puede ser de ayuda para la evaluación de los datos dado que podría evidenciar algunos tipos de valores que no son tan visibles o tan claros al usar otro tipo de algoritmos de aprendizaje y evaluación.

TÍTULO: Discovering key residues of dengue virus NS2b-NS3-protease: New binding sites for antiviral inhibitors design.

PALABRAS CLAVE: dengue virus, DENV NS2b-NS3 protease, Bindability sites, Protease inhibitor, Computational alanine scanning mutagénesis, Machine learning, Multilayer perceptrón**.**

FUENTE: Aguilera-Pesantes, D., Robayo, L. E., Méndez, P. E., Mollocana, D., Marrero-Ponce, Y., Torres, F. J., & Méndez, M. A. (2017). Discovering key residues of dengue virus NS2b-NS3-protease: New binding sites for antiviral inhibitors design. Biochemical and Biophysical Research Communications.

RESUMEN: The NS2B-NS3 protease is essential for the Dengue Virus (DENV) replication process. This complex constitutes a target for efficient antiviral discovery because a drug could inhibit the viral polyprotein processing. Furthermore, since the protease is highly conserved between the four Dengue virus serotypes, it is probable that a drug would be equally effective against all of them. In this article, a strategy is reported that allowed us to identify influential residues on the function of the Dengue NS2b-NS3 Protease. Moreover, this is a strategy that could be applied to virtually any protein for the search of alternative influential residues, and for non-competitive inhibitor development. First, we incorporated several features derived from computational alanine scanning mutagenesis, sequence, structure conservation, and other structure-based characteristics. Second, these features were used as variables to obtain a multilayer perceptron model to identify defined groups (clusters) of key residues as possible candidate pockets for binding sites of new leads on the DENV protease. The identified residues included: i) amino acids close to the beta sheet-loop-beta sheet known to be

important in its closed conformation for NS2b ii) residues close to the active site, iii) several residues evenly spread on the NS2b-NS3 contact surface, and iv) some inner residues most likely related to the overall stability of the protease. In addition, we found concordance on our list of residues with previously identified amino acids part of a highly conserved peptide studied for vaccine development.

APORTE A LA TESIS: Intentar descubrir cómo atacar el virus desde su ADN, para el control del dengue en sus diferentes serotipos y proteger a la humanidad de este tipo de enfermedad, se conoce que el problema real no es del todo los vectores que transmiten a enfermedad sino el virus que transportan, si se logra controlar el virus con un tipo de antiviral, la enfermedad casi que estaría en el olvido, pero reconocer cómo identificar, la maner de atacarla es la tarea difícil, con la evaluación de un modelo como el perceptrón multicapa para identificar los clusters o grupos definidos de residuos clave de derivaciones de la proteasa en el virus en sus serotipos definidos que son coincidentes en este componente y que es esencial para la replicación del virus dengue, es otra alternativa como referencia para identificar los factores que influyen en la replicación y expansión del virus o del vector.

TÍTULO: A Supervised Learning Process to Validate Online Disease Reports for Use in Predictive Models.

PALABRAS CLAVE: dengue**,** supervised learning, predictive models, Big data, socioeconomic factors.

FUENTE: Patching, H. M., Hudson, L. M., Cooke, W., Garcia, A. J., Hay, S. I., Roberts, M., & Moyes, C. L. (2015). A supervised learning process to validate online disease reports for use in predictive models. Big data, 3(4), 230-237.

RESUMEN: Pathogen distribution models that predict spatial variation in disease occurrence require data from a large number of geographic locations to generate disease risk maps. Traditionally, this process has used data from public health reporting systems; however, using online reports of new infections could speed up the process dramatically. Data from both public health systems and online sources must be validated before they can be used, but no mechanisms exist to validate data from online media reports. We have developed a supervised learning process to validate geolocated disease outbreak data in a timely manner. The process uses three input features, the data source and two metrics derived from the location of each disease occurrence. The location of disease occurrence provides information on the probability of disease occurrence at that location based on environmental and socioeconomic factors and the distance within or outside the current known disease extent. The process also uses validation scores, generated by disease experts who review a subset of the data, to build a training data set. The aim of the supervised learning process is to generate validation scores that can be used as weights going into the pathogen distribution model. After analyzing the three input features and testing the performance of alternative processes, we selected a cascade of ensembles comprising logistic regressors. Parameter values for the training data subset size, number of predictors, and number of layers in the cascade were tested before the process was deployed. The final configuration was tested using data for two contrasting

diseases (dengue and cholera), and 66%–79% of data points were assigned a validation score. The remaining data points are scored by the experts, and the results inform the training data set for the next set of predictors, as well as going to the pathogen distribution model. The new supervised learning process has been implemented within our live site and is being used to validate the data that our system uses to produce updated predictive disease maps on a weekly basis.

APORTE A LA TESIS: La integración de variables como los factores socioeconómicos de un sitio, así como la distancia que tiene una población contra un foco de infección, es un aporte valioso en una crisis de salud pública, usar los algoritmos de machine learning para proponer un mapa de riesgo de la enfermedad que puede ser válido tanto para el dengue como extrapolable para otras enfermedades como el cólera, la predicción de los brotes para su posterior creación de mapas de alertas de riesgo bajo la modalidad de aprendizaje supervisado, es una de las mezclas que permiten avanzar en la evaluación, control y erradicación del vector en las zonas de influencia, y también en nuevas locaciones donde se pueda presentar.

TÍTULO: Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM).

PALABRAS CLAVE: dengue, machine learning, Gaussian radial basis function, Raman spectroscopy, polynomial function, linear function, support vector machine.

FUENTE: Khan, S., Ullah, R., Khan, A., Wahab, N., Bilal, M., & Ahmed, M. (2016). Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). Biomedical optics express, 7(6), 2249-2256.

RESUMEN: The current study presents the use of Raman spectroscopy combined with support vector machine (SVM) for the classification of dengue suspected human blood sera. Raman spectra for 84 clinically dengue suspected patients acquired from Holy Family Hospital, Rawalpindi, Pakistan, have been used in this study. The spectral differences between dengue positive and normal sera have been exploited by using effective machine learning techniques. In this regard, SVM models built on the basis of three different kernel functions including Gaussian radial basis function (RBF), polynomial function and linear function have been employed to classify the human blood sera based on features obtained from Raman Spectra.The classification model have been evaluated with the 10-fold cross validation method. In the present study, the best performance has been achieved for the polynomial kernel of order 1. A diagnostic accuracy of about 85% with the precision of 90%, sensitivity of 73% and specificity of 93% has been achieved under these conditions.

APORTE A LA TESIS: La inclusión de campos como la microbiología con especial interés en el desarrollo de una manera de combatir los virus y en este caso el dengue, han llevado a la necesidad de utilizar técnicas de machine learning para sus estudios donde contrastan con otros análisis donde se usan por ejemplo SVM, este estudio adiciona técnicas como la función polinomial y la función lineal para la clasificación,

donde el campo de la biofotonica actualmente usa en gran parte de sus estudios las técnicas de machine learning para encontrar patrones que permitan entender los fenómenos que afectan a los estudios de ese campo, tener este tipo de algoritmos como fuente alterna de identificación puede ser provechoso para el análisis de los datos de esta tesis.

TÍTULO: Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machineand the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas.

PALABRAS CLAVE: machine learning, support vector machine, Aedes aegypti, forecasting, morbity, dengue.

FUENTE: Kesorn, K., Ongruk, P., Chompoosri, J., Phumee, A., Thavara, U., Tawatsin, A., & Siriyasatien, P. (2015). Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. PloS one, 10(5), e0125049.

RESUMEN: In the past few decades, several researchers have proposed highly accurate prediction models that have typically relied on climate parameters. However, climate factors can be unreliable and can lower the effectiveness of prediction when they are applied in locations where climate factors do not differ significantly. The purpose of this study was to improve a dengue surveillance system in areas with similar climate by

exploiting the infection rate in the Aedes aegypti mosquito and using the support vector machine (SVM) technique for forecasting the dengue morbidity rate.

APORTE A LA TESIS: La técnica metodológica que se propone en el estudio que se compone de integración de datos, construcción del modelo y evaluación del modelo es un marco interesante como referencia para la manera en que se puede llevar a cabo el análisis del estudio de los datos, consecuentemente refuerza la utilización del algoritmo de SVM y la introducción del núcleo de función de base radial RBF en el SVM que para este tipo de situaciones es muy utilizado y será clave para la elección de los algoritmos a usar, también el uso de factores como la sensibilidad, la especificidad y la precisión se usaron para evaluar la efectividad del modelo propuesto y se obtuvo una buena respuesta con relación a la predicción de la tasa de morbilidad en una zona que es similar a Colombia en algunos aspectos socio-demográficos y ambientales que pueden ser un buen conjunto de comparación en la obtención y discusión de resultados.

TÍTULO: A structured approach to predictive modeling of a two-class problem using multidimensional data sets.

PALABRAS CLAVE: Supervised learning, Classification, Data exploration, Machine learning, Data mining.

FUENTE: Spratt, H., Ju, H., & Brasier, A. R. (2013). A structured approach to predictive modeling of a two-class problem using multidimensional data sets. Methods, 61(1), 73-85.

RESUMEN: Biological experiments in the post-genome era can generate a staggering amount of complex data that challenges experimentalists to extract meaningful information. Increasingly, the success of an appropriately controlled experiment relies on a robust data analysis pipeline. In this paper, we present a structured approach to the analysis of multidimensional data that relies on a close, two-way communication between the bioinformatician and experimentalist. A sequential approach employing data exploration (visualization, graphical and analytical study), pre-processing, feature reduction and supervised classification using machine learning is presented. This standardized approach is illustrated by an example from a proteomic data analysis that has been used to predict the risk of infectious disease outcome. Strategies for model selection and post hoc model diagnostics are presented and applied to the case illustration. We discuss some of the practical lessons we have learned applying supervised classification to multidimensional data sets, one of which is the importance of feature reduction in achieving optimal modeling performance.

APORTE A LA TESIS: El enfoque para interpretación de información como el análisis multidimensional de datos, es otra alternativa útil dada la cantidad de variables que posee el conjunto de datos a evaluar y con el fin de poder sacar mejor partido a la información ya que este tiene un enfoque bidireccional y de comunicación cercana entre los

implicados, planteando un conjunto de hitos a seguir como un enfoque secuencial que emplea exploración de datos (visualización, estudio gráfico y analítico), preprocesamiento, reducción de características y clasificación supervisada mediante machine learning, técnicas interesantes como el uso de la validación cruzada y la tasa de descubrimiento falso, entre otras. Esta tasa de descubrimiento falso es bastante común en muchos de los diagnósticos que se creen positivos y luego se confirma que el diagnóstico era erróneo, esto genera un reporte de casos mal informado que puede generar por ejemplo, el mal uso de los recursos económicos y de talento humano hacia una zona. También es importante entender las estrategias que se utilizaron para la selección del modelo y el posterior diagnóstico de este, ya que aplicar una metodología adecuada para el tratamiento de datos es muy relevante a la hora de sacar las conclusiones del estudio.

TÍTULO: Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level.

PALABRAS CLAVE: dengue, forecasting, virus, machine learning, social networks, infection, Aedes, aegyti

FUENTE: de Almeida Marques-Toledo, C., Degener, C. M., Vinhal, L., Coelho, G., Meira, W., Codeço, C. T., & Teixeira, M. M. (2017). Dengue prediction by the web: tweets are a useful tool for estimating and forecasting dengue at country and city level. PLoS neglected tropical diseases, 11(7), e0005729.

RESUMEN: Infectious diseases are a leading threat to public health. Accurate and timely monitoring of disease risk and progress can reduce their impact. Mentioning a disease in social networks is correlated with physician visits by patients, and can be used to estimate disease activity. Dengue is the fastest growing mosquito-borne viral disease, with an estimated annual incidence of 390 million infections, of which 96 million manifest clinically. Dengue burden is likely to increase in the future owing to trends toward increased urbanization, scarce water supplies and, possibly, environmental change. The epidemiological dynamic of Dengue is complex and difficult to predict, partly due to costly and slow surveillance systems.

APORTE A LA TESIS: Utilizar tweets con palabras clave para identificar la incidencia del dengue en Brasil y tener una relación entre estos, es bastante fuera de lo común, utilizar la información que publican las personas que tienen el virus o que conocen a alguien que lo padece y poder predecir brotes en zonas es un modelo muy interesante para tener en cuenta dado que las personas publican masivamente a cada instante un gran volumen de información con los sucesos que están viviendo y mezclar temas como los índices sociodemográficos para predecir brotes de epidemias de dengue es realmente notable y útil ya que procesar la información publicada hace un día por miles de twitteros o personas de las redes sociales puede ayudar a identificar con mayor agilidad próximos brotes posibles, comparado con esperar que la entidad oficial local o nacional publique los resultados.

TÍTULO: Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción.

PALABRAS CLAVE: Bayes, Redes Bayesianas. Aprendizaje por Inducción. Clasificación, Sistemas inteligentes híbridos.

FUENTE: Felgaer, P. (2004). Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción. Reportes Técnicos en Ingeniería del Software, 6(2), 64-69.

RESUMEN: Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística; son utilizadas para proveer: una forma compacta de representar el conocimiento y métodos flexibles de razonamiento. El obtener una red bayesiana a partir de datos es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico. En este trabajo se define un método de aprendizaje automático que optimiza las redes bayesianas aplicadas a clasificación, mediante la utilización de un método de aprendizaje híbrido que combina las ventajas de las técnicas de inducción de los árboles de decisión (TDIDT - C4.5) con las de las redes bayesianas. El método resultante es aplicado a la predicción en dominios de salud.

APORTE A LA TESIS: Las técnicas de clasificación basadas en grafos permiten explicar algún tipo de fenómenos, en algunos casos de mejor manera, dado que muestra la dependencia de una variable con respecto a otras, tener una noción de aprendizaje de dos etapas: el aprendizaje estructural y el aprendizaje paramétrico, puede ser favorable

para la clasificación de los factores que más afectan a las zonas o poblaciones infestadas por el virus del dengue, tener una experimentación híbrida entre árboles de decisión y redes bayesianas ha dado buen resultado para la predicción en dominios de salud, aunque no se analiza el dengue podrían dar buena respuesta a los datos que se serán objetos de estudio.

TÍTULO: A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue.

PALABRAS CLAVE: Data Mining, Medical data, Machine learning algorithms, Diagnosis, Arbovirus.

FUENTE: Fathima, A. S., Manimegalai, D., & Hundewale, N. (2011). A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. IJCSI International Journal of Computer Science Issues, 8(6), 322-328.

RESUMEN: Chikungunya (CHIK) virus, similar to Dengue pose a serious threat in Tropics, because of the year-round presence of Aedes mosquito vectors. The use of machine learning techniques and data mining algorithms have taken a great role in the diagnosis and prognosis of many health diseases. But a very few work has been initialized in this arboviral medical informatics. Our focus is to observe clinical and physical diagnosis of chikungunya viral fever patients and its comparison with dengue viral fever. Our project aims to integrate different sources of information and to discover patterns of diagnosis, for predicting the viral infected patients and their results. The scope is mainly

in the classification problem of these often confused arboviral infections. This study paper summarizes various review and technical articles on arboviral diagnosis and prognosis. In this paper we present an overview of the current research being carried out using the data mining techniques to enhance the arboviral disease diagnosis and prognosis. This paper is not intended to provide a comprehensive overview of medical data mining but rather describes some areas which seem to be important from our point of view for applying machine learning in medical diagnosis for our real viral dataset.

APORTE A LA TESIS: Usar la minería de datos y algoritmos de machine learning en el proceso de descubrir los patrones que permitan predecir datos, ampliar el espectro de algoritmos de análisis como la utilización del algoritmo KDD, las redes neuronales, métodos como los de Fuzzy Sets y Rough Sets, y el concepto de sumarización para usar su metodología que utiliza diferentes tipos de algoritmos para evaluar un set de datos y consiste en once pasos con los que se busca intentar dar el mejor resultado, son puntos de vista alternos a la manera como se puede abordar el problema en su fase de análisis que permita obtener un mejor resultado.

TÍTULO: decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore.

PALABRAS CLAVE: dengue, decision tree algorithm, Singapore. Demographic, clinical, dengue haemorrhagic fever.

FUENTE: Lee, V. J., Lye, D. C., Sun, Y., & Leo, Y. S. (2009). Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore. Tropical Medicine & International Health, 14(9), 1154-1159.

RESUMEN: Differentiating dengue fever (DF) from more severe forms of dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS) in the early phases of illness are clinically challenging. Three studies of children (Teeraratkul et al. 1990; Kalayanarooj et al. 1997; Carlos et al. 2005) and one study of adults (Lee et al. 2006) attempted to compare DF and DHF patients to derive predictors of DHF, but none resulted in a clinically useful tool to assist clinicians in recommending hospitalization for patients with dengue in the early febrile phase of their illness. In Singapore, about 80% of notified adult dengue cases were hospitalized from 2000 to 2005 despite low rates of DHF (1.8–2.8%) (Lye et al. 2008). An easy-to-use tool for clinicians is therefore necessary to reduce unnecessary dengue admissions and to focus on the management of potential complicated cases.

APORTE A LA TESIS: Este trabajo de investigación permite tener una alternativa adicional para clasificar la prioridad de atención de las personas en una posible epidemia con el algoritmo del árbol de decisión, con esto se podría plantear un análisis que permita determinar que debido a la zona donde las personas se encuentren y sean infectadas, sea posible clasificarlo previamente y desviar a los infectados con el virus hacia los centros donde reciban la atención que necesiten y así intentar disminuir la tasa de mortalidad de las personas.

TÍTULO: Classification of dengue fever patients based on gene expression data using support vector machines.

PALABRAS CLAVE: support vector machines, machine learning, dengue, Classification, dengue fever (DF) to dengue haemorrhagic fever (DHF).

FUENTE: Gomes, A. L. V., Wee, L. J., Khan, A. M., Gil, L. H., Marques Jr, E. T., Calzavara-Silva, C. E., & Tan, T. W. (2010). Classification of dengue fever patients based on gene expression data using support vector machines. PloS one, 5(6), e11267.

RESUMEN: Symptomatic infection by dengue virus (DENV) can range from dengue fever (DF) to dengue haemorrhagic fever (DHF), however, the determinants of DF or DHF progression are not completely understood. It is hypothesised that host innate immune response factors are involved in modulating the disease outcome and the expression levels of genes involved in this response could be used as early prognostic markers for disease severity.

APORTE A LA TESIS: Con el planteamiento del uso del algoritmo de máquina de vectores de soporte planteado en esta investigación para clasificar las personas infectadas con las variaciones de dengue y dengue fuerte, permite entender la importancia de los genes involucrados, esto podría permitir dar una nueva separación de los diferentes tipos de casos que serán analizados y poder identificar zonas que tengan un mayor foco de mayor peligro para las personas.

TÍTULO: Chikungunya, Dengue, And Zika In Cali, Colombia: Epidemiological And Geospatial Analyses.

PALABRAS CLAVE: Dengue, chikungunya, and Zika, environmental, demographic, temporal, climate variables, geographical aspects.

FUENTE: Krystosik, A. R. (2016). CHIKUNGUNYA, DENGUE, AND ZIKA IN CALI, COLOMBIA: EPIDEMIOLOGICAL AND GEOSPATIAL ANALYSES (Doctoral dissertation, Kent State University).

RESUMEN: Dengue, chikungunya, and Zika are vector-borne diseases of global health concern. Cali, Colombia experienced hypoendemic dengue and outbreaks of Zika and chikungunya between October 2014 and April 2016. The epidemiological and geographical aspects of these diseases were investigated using classical epidemiological measures, geographically weighted regression modeling, and spatial video geonarratives. These diseases were found to be spatially clustered and related to key environmental, demographic, temporal, and climate variables at neighborhood and sub-neighborhood levels. These findings have implications for public health policy and vector control in Cali, Colombia.

APORTE A LA TESIS: Los modelos geográficos que se plantean en esta tesis doctoral usando modelos de regresión ponderados muestra que las enfermedades que se transmiten por vectores regularmente se agrupan espacialmente y tienen fuerte relación

ambiental, demográfica y climáticas entre otras, esto permite tener una perspectiva con las zonas que se deben atender y vigilar con más atención para evitar fuertes epidemias en tiempos futuros.

TÍTULO: Kernel-Based Machine Learning Models for the Prediction of Dengue and Chikungunya Morbidity in Colombia.

PALABRAS CLAVE: Machine learning, Forecasting, Dengue, Chikungunya, Kernel Ridge regression, Gaussian Processes**.**

FUENTE: Caicedo-Torres, W., Montes-Grajales, D., Miranda-Castro, W., Fennix-Agudelo, M., & Agudelo-Herrera, N. (2017, September). Kernel-Based Machine Learning Models for the Prediction of Dengue and Chikungunya Morbidity in Colombia. In Colombian Conference on Computing (pp. 472-484). Springer, Cham.

RESUMEN: Dengue and Chikungunya fever are two viral diseases of great public health concern in Colombia and other tropical countries as they are both transmitted by Aedes mosquitoes, which are endemic to this area. In recent years, there have been unprecedented outbreaks of these infections. Therefore, the development of computational models to forecast the number of cases based on available epidemiological data would benefit public surveillance health systems to take effective actions regarding the prevention and mitigation of these events. In this work, we present the application of machine learning algorithms to predict the morbidity dynamics of

dengue and chikungunya in Colombia using time-series-forecasting methods. Available weekly incidence for dengue (2007–2016) and chikungunya (2014–2016) from the National Health Institute of Colombia was gathered and employed as input to generate and validate the models. Kernel Ridge Regression and Gaussian Processes were used at forecasting the number of cases of both diseases considering horizons of one and four weeks. In order to assess the performance of the algorithms, rolling-origin cross-validation was carried out, and the mean absolute percentage errors (MAPE), mean absolute errors (MAE), R2 and the percentages of explained variance calculated for each model. Kernel Ridge regression with one-step ahead horizon was found to be superior to other models in forecasting both dengue and chikungunya number of cases per week. However, the power of prediction for dengue incidence was higher as there is more epidemiological data available for this disease compared to chikungunya. The results are promising and urge further research and development to achieve a tool which could be used by public health officials to manage more adequately the epidemiological dynamics of these diseases.

APORTE A LA TESIS: El documento muestra que se pueden utilizar algoritmos de machine learning como por ejemplo, los métodos de pronóstico de series de tiempo, para la predicción de la morbilidad de los vectores que transporten el virus del dengue en una región, también se hizo uso de otros métodos como la Regresión de Kernel Ridge y los Procesos Gaussianos con el fin de pronosticar la cantidad de casos de las enfermedades en un periodo de tiempo entre una y cuatro semanas. Lo que favorece la perspectiva de este trabajo debido a que los datos están clasificados por años.

TÍTULO: Zika virus disease, microcephaly and Guillain-Barré syndrome in Colombia: epidemiological situation during 21 months of the Zika virus outbreak, 2015–2017.

PALABRAS CLAVE: Infectious diseases vectors, Epidemiology, Culicidae, Morbidity, Nervous system congenital abnormalities, Education, Public health profesional, Population.

FUENTE: Arrieta, G., Caicedo-Castro, I., Oviedo-Pastrana, M., Méndez, N., & Mattar, S. (2017). Zika virus disease, microcephaly and Guillain-Barré syndrome in Colombia: epidemiological situation during 21 months of the Zika virus outbreak, 2015–2017. Archives of Public Health, 75(1), 65.

RESUMEN: The Zika virus disease (ZVD) has had a huge impact on public health in Colombia for the numbers of people affected and the presentation of Guillain-Barre syndrome (GBS) and microcephaly cases associated to ZVD.

APORTE A LA TESIS: El aporte del estudio para este documento es la metodología que se usa para tratar la información y la secuencia de actividades para el análisis de datos y de qué manera se puede sacar mejor partido de la información cuando se usan datos de varios periodos de tiempo.

TÍTULO: Estudios sobre dengue experiencias y perspectivas.

PALABRAS CLAVE: dengue, america, virus, togaviridae, OMS, flavivirus.

FUENTE: Valdes, M. G., & Valdes, M. G. (2012). Estudios sobre dengue experiencias y perspectivas (No. 61 610). e-libro, Corp.

RESUMEN: Con casi doscientos años de presencia, el Dengue, ha sido reportado por catorce países en América tras su primer brote importante, ocurrido en Cuba en 1981. Las estadísticas de la Organización Mundial de la Salud refieren que después de la década de los noventas, esta enfermedad, trasmitida por el mosquito Aedes aegipty, y producido por el virus Togaviridae, subgénero Flavivirus, ha emergido y evolucionado con fuerza, con el consiguiente riesgo para aquellas poblaciones susceptibles de adquirirla. Poblaciones que actualmente se cuentan en más de cien países con reportes estadísticos de cincuenta millones de casos aproximadamente en todo el mundo.

APORTE A LA TESIS: Este documento nos permite reforzar la parte teórica de la historia del dengue en las Américas, el vector que lo transmite, la importancia clínica y epidemiológica del dengue, acompañados de unos modelos de diagnóstico del dengue con técnicas de machine learning que ayudan en la selección de los algoritmos que más se pueden adecuar para el análisis de la información de este estudio.

TÍTULO: Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia.

PALABRAS CLAVE: dengue, Colombia, vectores, Aedes, epidemia**.**

FUENTE: Padilla, J. C., Rojas, D. P., & Gómez, R. S. (2012). Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia. Guías de Impresión Ltda.

RESUMEN: A pesar de existir abundante información relacionada con el comportamiento histórico del dengue en el país, son pocas las publicaciones sobre la sistematización, el análisis y la interpretación integral de la misma. Esta fue, tal vez, una de las principales motivaciones que tuvieron los autores para pensar, crear y desarrollar una propuesta que fuera novedosa, sencilla y útil para todos los usuarios interesados en el tema del dengue. Precisamente, la presente obra pretende, no sólo llenar ese vacío conceptual, sino contribuir a actualizar, consolidar y enriquecer los elementos teóricos existentes sobre el tema en el país, contextualizar y comprender mejor las múltiples causas y las consecuencias de un problema tan dinámico y variable, y aportar información estructurada para la toma de decisiones. Seguramente, aportará una línea de base y una fuente de hipótesis para que investigadores, académicos, profesionales y otros interesados en el tema puedan plantear preguntas específicas de investigación sobre diversos tópicos del dengue que contribuyan al avance del conocimiento de la enfermedad en el país**.**

APORTE A LA TESIS: Como el documento narra la historia del dengue en Colombia utilizando la centralización de fuentes históricas en este libro, con esta información se tiene noción de cómo ha llegado, se ha establecido, ha desaparecido y reemergido el virus en la nación y la evolución de éste a lo largo de los años, brinda un fuerte fundamento teórico a la tesis en la parte histórica.

4. EXPLORACIÓN DE LA INFORMACIÓN

I. ORIGEN DE LOS DATOS

La información que se utilizará para llevar a cabo este análisis, pertenece al departamento de Risaralda, los datos son correspondientes al dengue clásico, al dengue fuerte y los casos que llegaron a la muerte por dengue, el periodo para el cual fue posible adquirir la información corresponde a una línea de tiempo de 7 años, comprendido entre el año 2008 hasta el año 2014. La entidad de donde proviene la información es el Sivigila quien es el Sistema de Vigilancia Pública en Salud de Colombia quien tiene como responsabilidad el proceso de observación y análisis objetivo, sistemático y constante de los eventos en salud, el cual sustenta la orientación, planificación, ejecución seguimiento y evaluación de la práctica de la salud pública (Ministerio de Salud S. d., 2018).

II. PREPARACIÓN DE LOS DATOS

El archivo original de donde provienen los datos para esta evaluación contiene un total de 817.777 registros que cuentan con múltiples eventos de salud pública, y gran cantidad de características recolectadas, donde se tomará en cuenta solo la información referente a los eventos 210, 220 y 580, que son pertenecientes al dengue clásico, al dengue fuerte y mortalidad por dengue correspondientemente, las columnas seleccionadas para el análisis fueron: tipo de evento, ocupación, tipo de seguridad social, etnia, sexo, municipio y dirección de residencia, hospital donde fue atendido, municipio de procedencia, tipo de caso, área del caso, año, mes, semana y la edad, estos datos comprenden un total de 18507 registros.

III. REVISIÓN INICIAL DE LOS DATOS

Con un primer análisis a los datos que fueron separados inicialmente y que corresponden solo a los tres eventos relacionados con el dengue, iniciando este proceso se evidenció que en algunos de los casos del archivo se cuenta con registros donde la información es redundante, mal ingresada, está incompleta o simplemente no está diligenciada en algunos de los campos que contiene, estos registros incompletos han alcanzado un total del 87.93% (16.273 registros) del total de la muestra, por eso se hizo necesario evaluar registro por registro para escoger cuáles de estos deben ser estudiados y corregidos ya que cada registro del archivo es necesario para tomar en cuenta en el análisis global de la muestra y así poder interpretar de mejor manera el comportamiento del fenómeno en la población que está siendo afectada por este flagelo.

IV. DEPURACIÓN DE LA INFORMACIÓN

Al terminar la fase de exploración inicial de la información, fue necesario corregir todos los registros incompletos en algunos casos teniendo que tomar el nombre de la persona desde el set de datos original, ya que no tenía diligenciado por ejemplo el sexo y asumiendo que con el nombre se conoce el género. Aprovechando este proceso se realizó una depuración de los datos con la creación de nuevos atributos basados en la información original generando un nuevo archivo donde la información que fue transcrita para el análisis contiene los campos: código del evento, nombre del evento, edad de la cual se creó una nueva columna correspondiente al grupo etario con rangos de 20 años, municipio de residencia, dirección de residencia gracias a esto fue posible adicionar una

nueva columna que indica el estrato social de la persona, año del caso, mes del caso, semana del caso, tipo de caso probable, confirmado por laboratorio o por nexo epidemiológico, seguridad social, la etnia a la que pertenece el afectado, centro de salud donde fue atendido, ocupación, sexo, tipo de área donde habita el infectado y si el paciente fue hospitalizado o no.

Posteriormente dado que la información, en especial del sitio de residencia en varios de los registros estaba parcialmente diligenciada o no tenía ninguna información y es de un valor muy importante en este análisis, fue necesario aproximarlo al centro de salud donde fue atendido. Consecuentemente se realizó una subdivisión en zonas más específicas, dado que municipios como Pereira, Dosquebradas y Santa Rosa de Cabal están distribuidos por comunas y corregimientos, entonces se creó una nueva característica que tiene en cuenta la información de este nuevo atributo con el fin de ubicar el caso de manera más específica en el análisis, está se depuró sobre la base de la información de residencia o sitio de atención hospitalaria del infectado, aclarando que no se dejó de lado ningún registro del resto de la información que corresponde a los demás municipios del set de datos, por eso se creó una columna donde se muestra la localización de los casos por comunas, dentro de los municipios antes mencionados, Los casos de la residencia fuera del departamento de Risaralda se clasificó como correspondiente a la comuna otros.

Como se crearon varias columnas nuevas con información que pueden ser consideradas relevante para el análisis general y particular, se mostrarán a continuación un conjunto

de tablas donde se muestra la simbología que se utilizó en cada caso donde se representan los datos de tipo numérico a valor nominal, o en las cuales se tenía un identificador por palabra a un símbolo, dado que en el análisis de información puede ser más viable para reconocerlos con nombres o símbolos más cortos que los originales al mostrarlos en los diagramas o gráficas que se obtengan.

La tabla 12 contiene las simbologías para los municipios que hacen parte de los registros y su significado en el set de datos y gráficas:

Tabla 12. Símbolos Municipios.

NÚMERO	MUNICIPIO	SÍMBOLO
1	PEREIRA	PE
2	DOSQUEBRADAS	DQ
3	LA VIRGINIA	LV
4	SANTA ROSA	SR
5	SANTUARIO	ST
6	MARSELLA	MS
7	BELEN DE UMBRÍA	BU
8	MISTRATO	MT
9	LA CELIA	LC
10	QUINCHIA	QC
11	GUATICA	GT
12	BALBOA	BB
13	APÍA	AP
14	PUEBLO RICO	PR
15	OTROS	OT

La tabla 13 contiene las simbologías para las comunas que hacen parte de los registros y su significado en el set de datos y gráficas:

Tabla 13. Comunas por municipios 1.

MUNICIPIO	NÚMERO	COMUNAS
PEREIRA	1	FERROCARRIL
PEREIRA	2	OLÍMPICA
PEREIRA	3	SAN JOAQUÍN
PEREIRA	4	CUBA
PEREIRA	5	CIUDADELA DEL CAFÉ
PEREIRA	6	EL OSO
PEREIRA	7	PERLA DEL OTÚN
PEREIRA	8	CONSOTA
PEREIRA	9	EL ROCÍO
PEREIRA	10	EL POBLADO
PEREIRA	11	EL JARDÍN
PEREIRA	12	SAN NICOLAS
PEREIRA	13	CENTRO
PEREIRA	14	RÍO OTÚN
PEREIRA	15	BOSTON
PEREIRA	16	UNIVERSIDAD
PEREIRA	17	VILLAVICENCIO
PEREIRA	18	ORIENTE
PEREIRA	19	VILLASANTANA
PEREIRA	20	ALTAGRACIA
PEREIRA	21	ARABIA
PEREIRA	22	CAIMALITO
PEREIRA	23	CERRITOS
PEREIRA	24	LA FLORIDA
PEREIRA	25	PUERTO CALDAS
PEREIRA	26	COMBIA
PEREIRA	28	LA BELLA
PEREIRA	29	LA ESTRELLA
PEREIRA	30	MORELIA
PEREIRA	31	TRIBUNAS
OTROS	65	OTROS

La tabla 14 contiene las simbologías para las comunas que hacen parte de los registros y su significado en el set de datos y gráficas:

Tabla 14. Comunas por municipios 2.

ÁREA	NÚMERO	COMUNAS
DOSQUEBRADAS	32	COMUNA 1
DOSQUEBRADAS	33	COMUNA 2
DOSQUEBRADAS	34	COMUNA 3
DOSQUEBRADAS	35	COMUNA 4
DOSQUEBRADAS	36	COMUNA 5
DOSQUEBRADAS	37	COMUNA 6
DOSQUEBRADAS	38	COMUNA 7
DOSQUEBRADAS	39	COMUNA 8
DOSQUEBRADAS	40	COMUNA 9
DOSQUEBRADAS	41	COMUNA 10
DOSQUEBRADAS	42	COMUNA 11
DOSQUEBRADAS	43	COMUNA 12
LA VIRGINIA	44	LA VIRGINIA
SANTA ROSA	45	COMUNA 1 LA HERMOSA
SANTA ROSA	46	COMUNA 2 SUR
SANTA ROSA	47	COMUNA 3 CENTRO SUR
SANTA ROSA	48	COMUNA 4 CENTRO NORTE
SANTA ROSA	49	COMUNA 5 NORTE
SANTA ROSA	50	CORREGIMIENTO EL ESPAÑOL
SANTA ROSA	51	CORREGIMIENTO SANTA BARBARA
SANTA ROSA	52	CORREGIMIENTO EL MANZANILLO
SANTA ROSA	53	CORREGIMIENTOS DEL SUR
SANTA ROSA	54	CORREGIMIENTO LA CAPILLA
SANTUARIO	55	SANTUARIO
MARSELLA	56	MARSELLA
BELÉN DE UMBRÍA	57	BELÉN DE UMBRÍA
MISTRATÓ	58	MISTRATÓ
LA CELIA	59	LA CELIA
QUINCHIA	60	QUINCHIA
GUATICA	61	GUATICA
BALBOA	62	BALBOA
APIA	63	APIA
PUEBLO RICO	64	PUEBLO RICO

La tabla 15 contiene las simbologías para las etnias contagiadas y su significado en el set de datos y gráficas.

Tabla 15. Símbolos de etnias.

NÚMERO	ETNIAS	SÍMBOLO
6	ÍNDIGENA	IN
5	ROM. GITANO	RG
4	RAIZAL	RA
3	PALENQUERO	PA
2	NEGRO MULATO	NM
1	OTRO	OT

La tabla 16 contiene las simbologías para los estratos involucrados y su significado en el set de datos y gráficas.

Tabla 16. Símbolos de estratos.

ESTRATO	SÍMBOLO
1	F
2	E
3	D
4	C
5	B
6	A

La tabla 17 contiene las simbologías para los tipos de seguridad social de los implicados y su significado en el set de datos y gráficas:

Tabla 17. Símbolos de seguridad social.

NÚMERO	SEGURIDAD SOCIAL	SÍMBOLO
6	EXCEPCIÓN	P
5	CONTRIBUTIVO	C
4	NO ASEGURADO	N
3	ESPECIAL	E
2	SUBSIDIADO	S
1	INDETERMINADO	I

La tabla 18 contiene las simbologías para los grupos etarios infectados y su significado en el set de datos y gráficas:

Tabla 18. Símbolos de grupo etario.

GRUPO_ETARIO	RANGO	SÍMBOLO
1	0-19	A
2	20-39	B
3	40-59	C
4	60-79	D
5	80-99	E
6	100-119	F

La tabla 19 contiene las simbologías para los tipos de confirmación de casos de las personas afectadas y su significado en el set de datos y gráficas:

Tabla 19. Símbolos de Tipo de Caso.

NÚMERO	CASO_TIPO	SÍMBOLO
1	Confirmado Laboratorio	CL
1	Nexo Epidemiológico	NE
0	Probable	PR

La tabla 20 contiene las simbologías para las áreas donde se detectaron los casos y su significado en el set de datos y gráficas:

Tabla 20. Símbolos de tipos de área de ocupación.

NÚMERO	ÁREA_TIPO	SÍMBOLO
1	Cabecera_Municipal	CM
2	Centro_Poblado	CP
3	Rural_Diperso	RD

La tabla 21 contiene las simbologías para conocer las profesiones de las personas con el virus, su nivel de ingreso aproximado con la profesión y su significado en el set de datos y gráficas:

Tabla 21. Símbolos de profesiones.

OCUPACIÓN	PROFESIOÓN	SÍMBOLO	INGRESOS	SÍMBOLO
1	poder ejecutivo	PE	1	A
2	profesionales universitarios	PU	2,3	B,C
0	fuerza pública	FP	3	C
3	técnicos y asistentes	TA	4,5	D,E
6	agricultura y agropecuarios	AG	5	E
4	empleos de oficina	EO	6	F
7	operarios, artesanos, construcción, mineros	AR	7,8	G,H
8	operarios madera, metales, minerales	OP	8,9	H,I
5	servicios y vendedores	SV	9,10	I,J
9	trabajadores no calificados	TN	10	J

V. ANÁLISIS PRELIMINAR DE LA INFORMACIÓN

En la primera aproximación a la evaluación de los datos se realizó un análisis con técnicas básicas de minería de datos sobre los dos tipos del set de datos usando el lenguaje de programación Python con los paquetes de análisis científico; este primer paso consistía en conocer como estaba distribuida la muestra. Otro de los temas a considerar en esta fase, es que no todos los casos de dengue fueron confirmados como positivos, así que algunos de estos casos son solo sospechas de que la persona tiene el virus, por esta razón se trabajará con un set de datos simplificado para los valores de los casos totalmente confirmados que reduce sustancialmente los registros de set de datos, pero de igual manera para la caracterización de las variables que se buscan en este

trabajo, se mantendrán los dos set de datos, tanto el que solo es casos confirmados, así como la mezcla de confirmados y probables. Con este hallazgo de casos confirmados los datos se redujeron a un total de 5.826 registros, los cuales representan un 31.48% del total de registros que es menos de la tercera parte del set de datos, Para efectos prácticos, de ahora en adelante, se mencionarán los casos totales reportados, como casos totales y los casos confirmados reportados, como casos confirmados o casos depurados, también se usarán ambos casos de set de datos para los análisis posteriores para efectos de evaluar la variación de los patrones en cada una de las muestras.

Tenemos 2 grupos, uno con la cantidad de casos que tiene la muestra como reportados con dengue, que corresponde al set de datos que contiene la totalidad de los registros y otro con la cantidad de casos que tiene la muestra como casos confirmados con dengue esos datos corresponden al set de datos que cuenta con toda la información, excepto aquellos que no fueron confirmados.

Es por esta razón que se tienen 2 grupos, lo que se puede observar en las siguientes ilustraciones 3 y 4.

Ilustración 3. Casos Totales Reportados de dengue entre 2008 y 2014.

Número de casos	No ponderados	18507
	Ponderados	18507

Ilustración 4. Casos Confirmados Reportados de dengue entre 2008 y 2014.

Número de casos	No ponderados	18507
	Ponderados	5826

Consecuentemente se revisa la cantidad de casos por municipio que existen en la muestra, el total de casos por municipio y el porcentaje de casos que cada municipio aporta al total de la muestra, para mejor entendimiento se inicia el proceso de tener resúmenes para los datos con todos los casos posibles de contagio por dengue, así como en los casos donde el virus fue confirmado como positivo, en las personas de la muestra como se puede observar en las siguientes ilustraciones 5 y 6.

Ilustración 5. Casos Totales Por Municipio.

	Valor	Recuento	Porcentaje
1	APIA	38	0,2%
2	BALBOA	59	0,3%
3	BELEN DE UMBRIA	266	1,4%
4	DOSQUEBRADAS	4395	23,7%
5	GUATICA	57	0,3%
6	LA CELIA	59	0,3%
7	LA VIRGINIA	909	4,9%
8	MARSELLA	255	1,4%
9	MISTRATO	152	0,8%
10	OTROS	413	2,2%
11	PEREIRA	10578	57,2%
12	PUEBLO RICO	137	0,7%
13	QUINCHIA	142	0,8%
14	SANTA ROSA	732	4,0%
15	SANTUARIO	315	1,7%

Ilustración 6. Casos Confirmados por Municipio.

	Valor	Recuento	Porcentaje
1	APIA	9	0,2%
2	BALBOA	8	0,1%
3	BELEN DE UMBRIA	77	1,3%
4	DOSQUEBRADAS	1462	25,1%
5	GUATICA	11	0,2%
6	LA CELIA	14	0,2%
7	LA VIRGINIA	256	4,4%
8	MARSELLA	63	1,1%
9	MISTRATO	31	0,5%
10	OTROS	115	2,0%
11	PEREIRA	3336	57,3%
12	PUEBLO RICO	44	0,8%
13	QUINCHIA	64	1,1%
14	SANTA ROSA	250	4,3%
15	SANTUARIO	86	1,5%

Como se puede apreciar en las ilustraciones anteriores, los municipios que más participación tienen en la muestra para el set de datos completo son: Pereira, Dosquebradas, La Virginia y Santa Rosa de Cabal con un total de 16.614 casos, en donde el municipio de Pereira es el mayor afectado con 10.578 casos en el periodo de evaluación del set de datos. Dosquebradas a su vez registró un total de 4.395 afectados, La Virginia alcanzó la cifra de 909 casos y santa rosa de cabal alcanzó los 732 casos. En el set de datos confirmados la tendencia inicia en Pereira como el mayor afectado con 3.336 casos en el periodo de evaluación del set de datos. Dosquebradas a su vez registro un total de 1.462 afectados, La Virginia alcanzo la cifra de 256 casos y Santa Rosa de Cabal llegó a los 250 casos.

Para entender mejor los datos de la ilustración anterior que es muy poco descriptiva ya que hace una evaluación global de los datos y no permite mostrar de alguna manera algún tipo de patrón visible, se evaluaron los datos por municipio y por cada año, para valorar si existe algún tipo de secuencia en los casos, por esto se muestra en las ilustraciones 7 y 8, la cantidad de casos por años en cada municipio del set de datos.

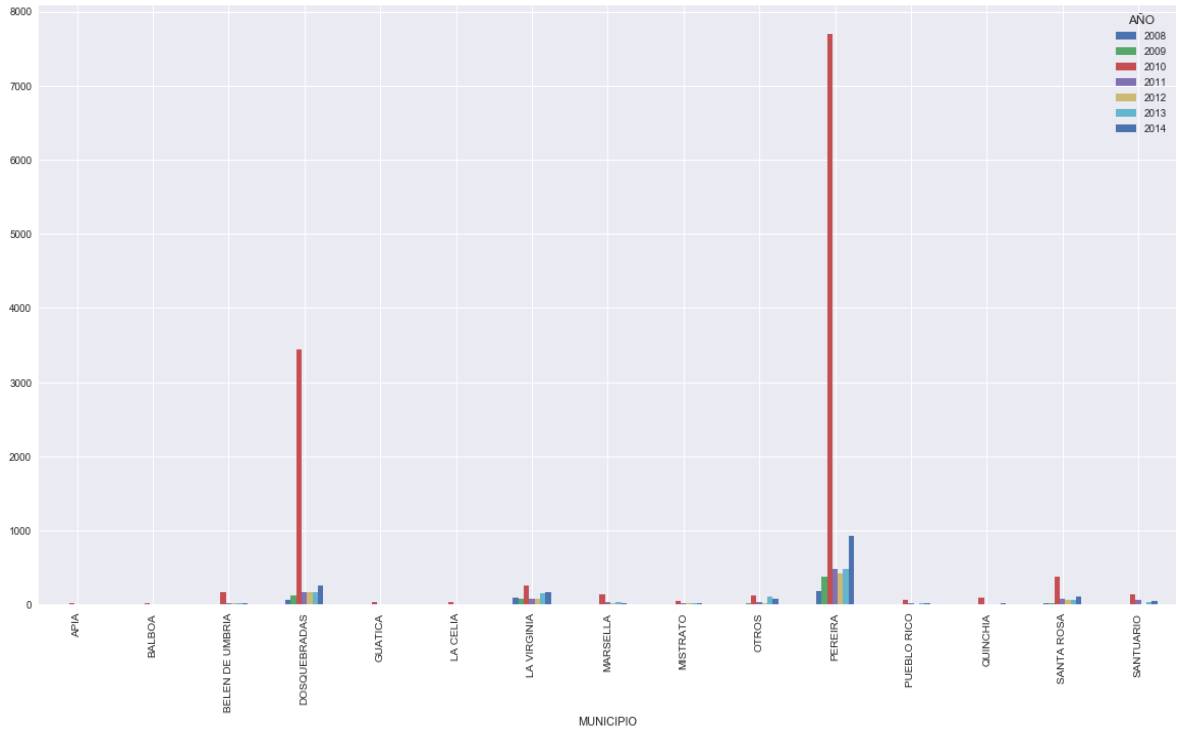
Ilustración 7. Casos Completos Anuales por Municipio.

	AÑOS							
MUNICIPIO	2008	2009	2010	2011	2012	2013	2014	Margen activo
APIA	1	4	21	4	1	1	6	38
BALBOA	0	1	22	4	8	13	11	59
BELEN DE UMBRIA	2	4	176	14	23	22	25	266
DOSQUEBRADAS	70	130	3441	163	167	170	254	4395
GUATICA	1	2	31	6	3	8	6	57
LA CELIA	2	2	34	5	8	6	2	59
LA VIRGINIA	92	75	266	74	75	153	174	909
MARSELLA	2	1	140	37	24	34	17	255
MISTRATO	8	12	57	15	15	27	18	152
OTROS	12	26	124	30	28	114	79	413
PEREIRA	189	371	7700	481	421	485	931	10578
PUEBLO RICO	3	5	64	21	7	14	23	137
QUINCHIA	7	2	89	7	6	12	19	142
SANTA ROSA	19	20	373	75	69	64	112	732
SANTUARIO	0	5	142	62	9	39	58	315
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 8. Casos Confirmados Anuales por Municipio.

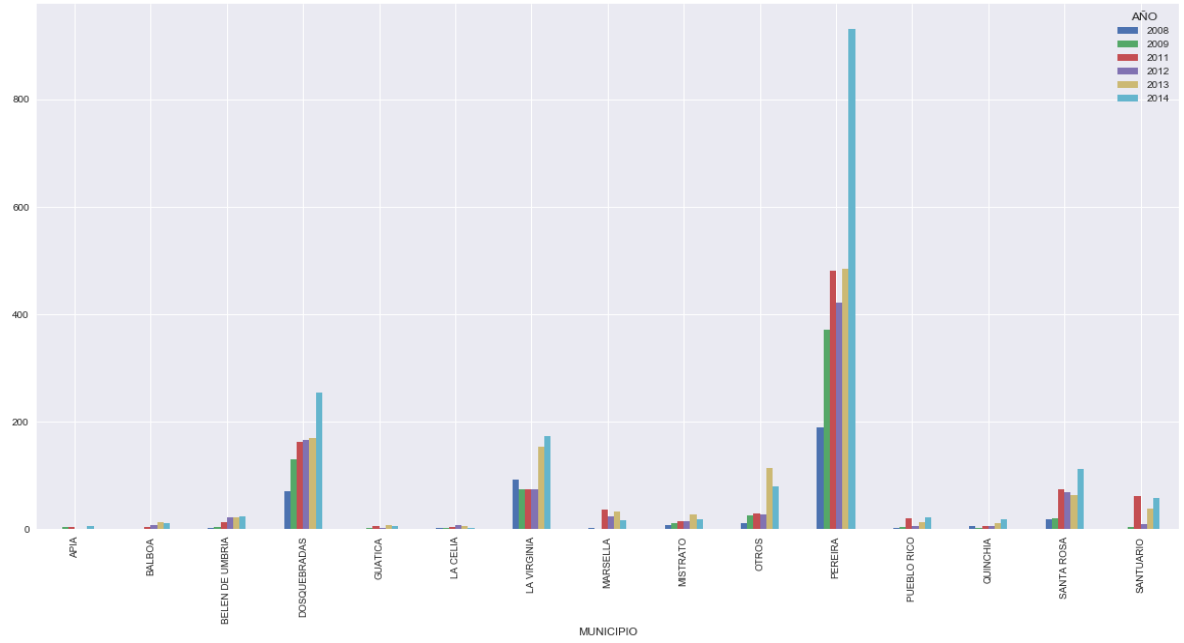
MUNICIPIO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
APIA	0	3	4	1	0	1	0	9
BALBOA	0	0	3	0	1	3	1	8
BELEN DE UMBRIA	2	4	54	2	3	7	5	77
DOSQUEBRADAS	25	83	1216	27	21	40	50	1462
GUATICA	1	1	6	1	0	1	1	11
LA CELIA	1	2	9	0	0	1	1	14
LA VIRGINIA	28	39	123	7	6	41	12	256
MARSELLA	1	1	41	6	2	10	2	63
MISTRATO	1	7	9	2	1	5	6	31
OTROS	4	14	40	8	4	20	25	115
PEREIRA	58	235	2672	69	40	95	167	3336
PUEBLO RICO	0	2	36	0	0	0	6	44
QUINCHIA	3	2	48	1	0	2	8	64
SANTA ROSA	10	13	170	23	6	12	16	250
SANTUARIO	0	0	61	8	1	8	8	86
Margen activo	134	406	4492	155	85	246	308	5826

Ilustración 9. Casos totales por Municipio y Año con mayor cantidad de casos 2008 -2014.



Como la ilustración 9 no facilita la visualización de los casos de dengue para todo el periodo evaluado, dado que el año 2010 fue un año con comportamientos atípicos, se construyó la gráfica de los demás años con la cual se visualiza de mejor manera los casos anuales por municipios en la ilustración 10.

Ilustración 10. Casos totales por Municipio y Año con mayor cantidad de casos 2008 -2009-2011-2012-2013-2014.



Como se puede percibir en las ilustraciones anteriores, entre los años 2008 a 2014 hay un crecimiento en casi todos los municipios al transcurrir los años o están dentro de un rango similar de casos entre años por municipio, a excepción del año 2010, donde claramente se nota que hubo un gran brote muy significativo comparado con los casos de los años anteriores, gracias a esta observación se entiende que no sería pertinente evaluar la información de todos los años como un mismo set de datos o un archivo único, dado que podría sesgar para unos propósitos de este trabajo el análisis de la información en cuanto a líneas de tiempo, pero para propósitos de otro tipo de variables como las espaciales, las sociales y económicas podría mostrar la tendencia en cuanto los casos comienzan a aumentar de manera desmesurada. Como el set de datos además de los

años, contiene los meses de los casos y también las semanas del brote, esta información se puede aprovechar para evaluar el comportamiento del virus en cada año y mes, también para el entrenamiento y la predicción de algún tipo de casos de dengue. Como Pereira, Dosquebradas y Santa Rosa, están divididas por comunas y cada comuna reporta su cantidad de casos, se percibe cual es el comportamiento del vector en cada una de las zonas dentro de cada municipio.

Ilustración 11. Casos Totales Anuales por Comunas de Pereira.

	AÑO							
COMUNA	2008	2009	2010	2011	2012	2013	2014	Margen activo
ALTAGRACIA	5	9	187	12	10	11	22	256
ARABIA	5	9	186	12	10	11	22	255
BOSTON	2	3	66	4	4	4	8	91
CAIMALITO	2	4	75	5	4	4	9	103
CENTRO	0	0	8	0	0	0	1	9
CERRITOS	4	7	154	10	8	9	19	211
CIUDADELA DEL CAFÉ	19	27	558	35	31	33	67	770
COMBIA	9	18	382	24	42	22	48	545
CONSOTA	1	3	59	4	3	3	7	80
CUBA	38	67	1548	96	85	123	185	2142
EL JARDIN	1	1	21	2	1	1	3	30
EL OSO	8	16	332	21	18	20	40	455
EL POBLADO	7	14	297	19	16	17	35	405
EL ROCIO	4	8	166	10	9	10	20	227
FERROCARRIL	8	16	331	21	18	19	40	453
LA BELLA	10	20	413	26	23	26	50	568
LA ESTRELLA	4	8	163	10	9	20	20	234
LA FLORIDA	1	1	25	2	1	2	3	35
MORELIA	5	11	220	14	12	13	27	302
OLIMPICA	1	3	56	3	3	3	7	76
ORIENTE	2	3	70	4	4	4	8	95
PERLA DEL OTUN	11	21	440	27	24	26	53	602
PUERTO CALDAS	3	6	118	7	6	7	15	162
RIO OTUN	4	8	171	11	6	10	0	210
SAN JOAQUIN	20	36	799	45	24	46	100	1070
SAN NICOLAS	0	6	128	8	7	0	15	164
TRIBUNAS	4	7	151	9	8	0	18	197
UNIVERSIDAD	1	2	34	2	2	0	4	45
VILLASANTANA	6	18	373	26	23	31	61	538
VILLAVICENCIO	4	19	169	12	10	10	24	248
Margen activo	189	371	7700	481	421	485	931	10578

En la ilustración 11 se aprecia que las comunas donde la incidencia del dengue es mayor, pertenecen a CUBA con un total de 2.142 casos y la otra zona que supera los mil casos es la comuna SAN JOAQUÍN con un reporte total de 1.070 infectados.

Ilustración 12. Casos Confirmados Anuales por Comunas de Pereira.

COMUNA	AÑO							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
ALTAGRACIA	1	8	85	2	0	1	3	100
ARABIA	2	7	71	2	1	1	3	87
BOSTON	0	3	22	0	0	0	2	27
CAIMALITO	0	3	32	0	0	0	2	37
CENTRO	0	0	4	0	0	0	0	4
CERRITOS	1	4	73	1	2	2	3	86
CIUDADELA DEL CAFÉ	6	21	175	6	2	7	9	226
COMBIA	5	9	147	1	7	5	7	181
CONSOTA	0	1	28	0	0	1	0	30
CUBA	10	49	483	14	9	29	42	636
EL JARDIN	0	0	6	1	1	0	0	8
EL OSO	1	9	102	4	0	8	7	131
EL POBLADO	0	10	121	2	1	0	6	140
EL ROCIO	1	5	63	1	0	0	1	71
FERROCARRIL	6	12	91	5	3	7	2	126
LA BELLA	2	11	156	1	0	2	13	185
LA ESTRELLA	3	7	70	4	0	2	2	88
LA FLORIDA	0	0	7	1	0	0	0	8
MORELIA	0	7	76	1	0	4	5	93
OLIMPICA	1	1	21	1	0	0	1	25
ORIENTE	0	3	16	0	0	1	1	21
PERLA DEL OTUN	4	13	143	2	3	3	8	176
PUERTO CALDAS	2	3	29	1	1	0	4	40
RIO OTUN	1	4	73	4	1	3	0	86
SAN JOAQUIN	7	18	255	5	2	10	18	315
SAN NICOLAS	0	3	54	0	1	0	3	61
TRIBUNAS	1	2	56	0	0	0	2	61
UNIVERSIDAD	0	1	21	1	0	0	0	23
VILLASANTANA	3	10	135	6	5	7	19	185
VILLAVICENCIO	1	11	57	3	1	2	4	79
Margen activo	58	235	2672	69	40	95	167	3336

En la ilustración 12, que contiene los casos de dengue confirmados, se observa que las comunas donde la incidencia del dengue es mayor, pertenecen a CUBA con un total de 636 casos y la otra zona que tiene alta cantidad de casos es la comuna SAN JOAQUÍN con un reporte total de 315 infectados. Ambas comunas son las líderes tanto para datos totales, así como para datos de casos confirmados.

Ilustración 13. Casos Totales Anuales por Comunas de Dosquebradas.

COMUNA	AÑO							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
COMUNA 1	2	4	95	5	5	5	7	123
COMUNA 10	3	6	171	8	8	8	13	217
COMUNA 11	3	5	124	6	7	6	9	160
COMUNA 12	3	5	124	6	6	7	9	160
COMUNA 2	2	4	112	5	5	6	8	142
COMUNA 3	3	6	167	8	8	8	12	212
COMUNA 4	2	3	75	4	4	4	6	98
COMUNA 5	1	3	73	3	4	4	5	93
COMUNA 6	7	14	363	17	18	18	27	464
COMUNA 7	6	11	298	14	14	15	22	380
COMUNA 8	32	58	1566	73	74	74	114	1991
COMUNA 9	6	11	273	14	14	15	22	355
Margen activo	70	130	3441	163	167	170	254	4395

En el municipio de Dosquebradas para los datos de casos totales, la ilustración 13 muestra que las zonas que reportan mayor cantidad de casos son la comuna 8 con 1.991 reportes, seguido de la comuna 6 con 464 reportes de enfermos por dengue y la comuna 7 con 380 afectados por el virus.

Ilustración 14. Casos Confirmados Anuales por Comunas de Dosquebradas.

COMUNA	AÑO							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
COMUNA 1	0	3	31	2	0	3	1	40
COMUNA 10	1	3	72	4	1	3	5	89
COMUNA 11	1	5	45	0	1	1	0	53
COMUNA 12	1	2	52	1	0	3	0	59
COMUNA 2	0	1	41	1	1	1	0	45
COMUNA 3	0	6	60	1	3	1	3	74
COMUNA 4	0	1	31	0	0	2	2	36
COMUNA 5	0	0	34	0	0	4	0	38
COMUNA 6	3	9	122	2	2	2	8	148
COMUNA 7	2	9	112	4	1	4	2	134
COMUNA 8	13	36	506	10	12	14	23	614
COMUNA 9	4	8	110	2	0	2	6	132
Margen activo	25	83	1216	27	21	40	50	1462

Para los datos de casos confirmados por dengue, la ilustración 14 muestra que la tendencia de mayor número de casos, se mantiene en las mismas zonas liderado por la comuna 8 con 614 ocurrencias, seguido por la comuna 6 con 148 reportes de enfermos

por dengue, la comuna 7 con 134 afectados y muy cerca la comuna 9 con 132 reportes del virus.

Ilustración 15. Casos Totales Anuales por Comunas de Santa Rosa.

COMUNA	AÑO							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
COMUNA 1 LA HERMOSA	4	4	71	14	13	12	21	139
COMUNA 2 SUR	3	3	53	11	10	9	16	105
COMUNA 3 CENTRO SUR	4	4	74	15	14	13	22	146
COMUNA 4 CENTRO NORTE	4	4	69	14	13	12	20	136
COMUNA 5 NORTE	3	3	65	13	12	10	20	126
CORREGIMIENTO EL ESPAÑOL	0	0	13	3	2	2	4	24
CORREGIMIENTO EL MANZANILLO	0	1	7	1	1	2	1	13
CORREGIMIENTO LA CAPILLA	0	0	9	2	2	2	3	18
CORREGIMIENTO SANTA BARBARA	1	1	10	2	2	2	4	22
CORREGIMIENTOS DEL SUR	0	0	2	0	0	0	1	3
Margen activo	19	20	373	75	69	64	112	732

Por último, se conoce con la ilustración 15 como son repartidos los casos de dengue en Santa Rosa para la totalidad de casos reportados, en donde las comunas 1 a la 5 reportan la mayor cantidad de casos en un valor muy cercano entre ellos y donde los corregimientos son afectados en una cantidad muy baja con respecto a las comunas.

Ilustración 16. Casos Confirmados Anuales por Comunas de Santa Rosa.

COMUNA	AÑO							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
COMUNA 1 LA HERMOSA	1	3	32	5	1	2	4	48
COMUNA 2 SUR	1	2	27	3	1	1	2	37
COMUNA 3 CENTRO SUR	4	4	37	3	2	2	3	55
COMUNA 4 CENTRO NORTE	3	1	25	6	0	3	2	40
COMUNA 5 NORTE	1	2	30	2	2	2	4	43
CORREGIMIENTO EL ESPAÑOL	0	0	4	2	0	0	0	6
CORREGIMIENTO EL MANZANILLO	0	1	6	0	0	0	1	8
CORREGIMIENTO LA CAPILLA	0	0	4	2	0	0	0	6
CORREGIMIENTO SANTA BARBARA	0	0	4	0	0	2	0	6
CORREGIMIENTOS DEL SUR	0	0	1	0	0	0	0	1
Margen activo	10	13	170	23	6	12	16	250

En los casos de dengue confirmados, la ilustración 16 muestra que se mantiene la tendencia en los reportes de infección del virus, además la similitud en cantidad de casos para las comunas y la poca participación de los corregimientos en los casos reportados de dengue para este municipio.

VI. CARACTERÍSTICAS DE LA INFORMACIÓN

Como parte de los objetivos de este trabajo es conocer qué tipo de variables influencian las enfermedades transmitidas por vectores, se hace necesario tomar en consideración cada una de las características que se aislaron en el set de datos para el análisis y así estudiar si estás pueden estar relacionadas con el flagelo, para dilucidar cómo defenderse de futuras posibilidades de contagios o emergencias en la zona que está siendo objeto de estudio y también en lo posible extrapolar el análisis si los datos tienen similitud con los datos de otras zonas del país.

GÉNERO DEL INFECTADO

Uno de los factores que podrían ser importantes para este análisis es conocer si el vector tiene preferencia en contagiar algún sexo en mayor cantidad, por lo cual en las siguientes ilustraciones 17, 18 y 19 se tendrá en cuenta esta hipótesis.

Ilustración 17. Casos globales por género y set de datos.

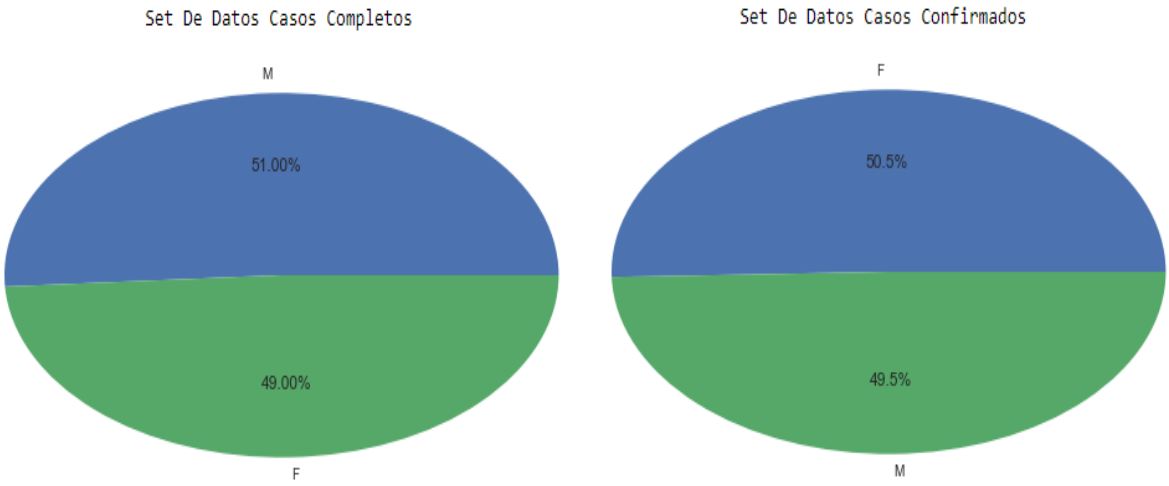


Ilustración 18. Casos Anuales Género Casos Completa.

GENERO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
F	208	316	6377	446	367	543	811	9068
M	200	344	6303	552	497	619	924	9439
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 19. Casos Anuales Género Casos Confirmados.

GENERO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
F	66	209	2288	80	36	119	142	2940
M	68	197	2204	75	49	127	166	2886
Margen activo	134	406	4492	155	85	246	308	5826

TIPOS DE ÁREAS DE LOS CASOS

Las zonas donde se desarrollan los casos pueden ser importantes para la identificación de los focos de infección más comunes, para este caso se cuenta con tres características, como son las cabeceras municipales, los centros poblados y los sitios rurales dispersos, estos se muestran en las siguientes ilustraciones 22, 21 y 22.

Ilustración 20. Tipos de áreas donde se confirmaron los casos.

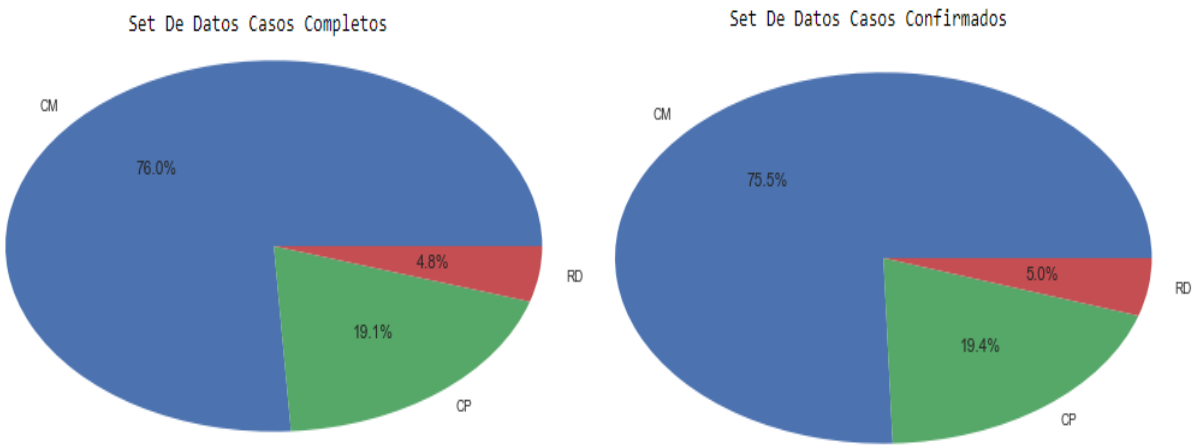


Ilustración 21. Cantidad de Casos por Área Casos Completas.

AREA	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
CM	313	489	9762	708	634	809	1267	13982
CP	75	137	2350	214	171	274	372	3593
RD	20	34	568	76	59	79	96	932
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 22. Cantidad de Casos por Área Casos Confirmados.

AREA	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
CM	100	308	3446	113	68	169	225	4429
CP	25	79	840	31	11	60	69	1115
RD	9	19	206	11	6	17	14	282
Margen activo	134	406	4492	155	85	246	308	5826

GRUPOS ETARIOS DE LOS CASOS

La edad que tiene una persona infectada es en buena medida un indicador de quiénes pueden ser los blancos más vulnerables para los mosquitos, entonces conocer si hay preferencia en este tema por parte del vector o si las infecciones no son referentes con las edades de las personas puede conducir a entender mejor los patrones de diseminación del vector en una población, se observa su comportamiento en las ilustraciones 23, 24 y 25.

Ilustración 23. Casos por grupo etario anuales.

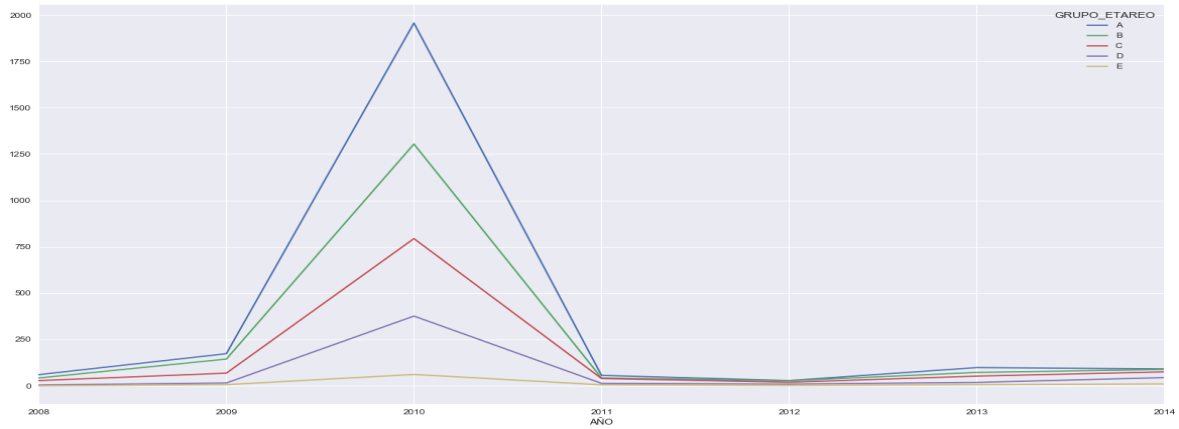


Ilustración 24. Grupos Etarios Datos Completos.

ETARIO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
A	167	272	5517	351	346	456	561	7670
B	136	244	3854	265	233	342	489	5563
C	86	104	2177	248	216	252	463	3546
D	19	27	994	101	59	75	176	1451
E	0	13	136	33	10	37	46	275
F	0	0	2	0	0	0	0	2
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 25. Grupos Etarios Datos Confirmados.

ETARIO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
A	60	173	1957	56	28	98	91	2463
B	42	144	1304	42	27	72	88	1719
C	28	68	794	39	19	52	75	1075
D	4	15	376	13	9	18	44	479
E	0	6	61	5	2	6	10	90
F	0	0	0	0	0	0	0	0
Margen activo	134	406	4492	155	85	246	308	5826

TIPOS DE ESTRATOS DE LOS CASOS

La estratificación de las personas que fueron afectadas por el virus es otro de los factores que puede ser distintivo a la hora de implementar algún tipo de estudio futuro para solucionar los brotes, con las siguientes ilustraciones 26, 27 y 28 se podrá revisar si es un factor incidente en la permanencia o ataque del vector.

Ilustración 26. Casos de Tipos de estratos por año.

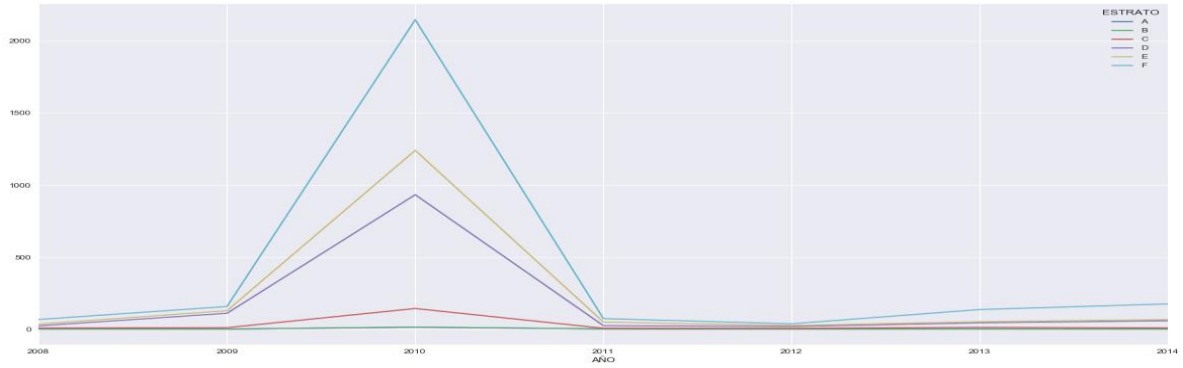


Ilustración 27. Casos de dengue por estrato, datos Completos.

ESTRATO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
A	1	1	36	5	3	4	8	58
B	2	1	21	9	5	14	9	61
C	14	17	375	48	31	52	80	617
D	73	163	2564	179	156	199	307	3641
E	116	193	3441	272	243	273	400	4938
F	202	285	6243	485	426	620	931	9192
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 28. Casos de dengue por estrato, datos Confirmados.

ESTRATO	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
A	1	0	14	1	0	2	0	18
B	1	1	13	2	0	1	1	19
C	7	10	143	6	5	12	8	191
D	23	110	933	24	18	44	58	1210
E	35	128	1241	49	25	51	66	1595
F	67	157	2148	73	37	136	175	2793
Margen activo	134	406	4492	155	85	246	308	5826

TIPOS DE SEGURIDAD SOCIAL DE LOS CASOS

La seguridad social de las personas puede ser de vital importancia al manifestarse un brote, dado que si no tienen acceso a la salud o su acceso es de mala calidad el cuadro clínico en caso de infección pude ser muy grave, ya que si no reciben los cuidados y atención necesaria es muy posible que el cuadro clínico se complique, en las siguientes ilustraciones 29, 30 y 31 se observa lo señalado.

Ilustración 29. Tipos de seguridad social por año donde se confirmaron los casos.

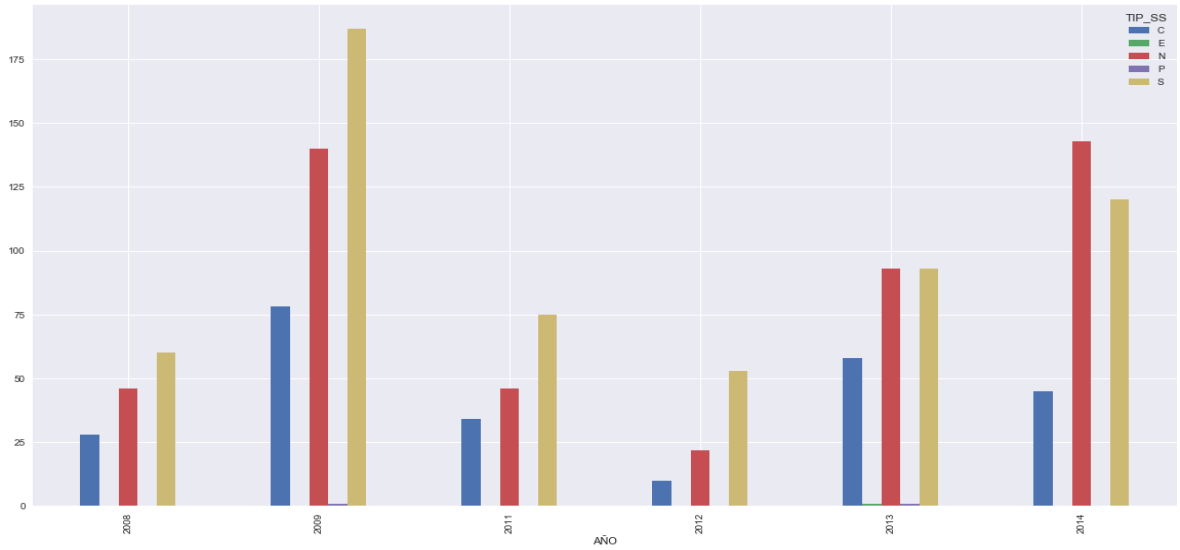


Ilustración 30. Casos de Seguridad Social datos Completos.

SS	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
C	75	121	2627	187	112	322	268	3712
E	0	0	1	0	0	1	0	2
N	150	245	4310	275	337	416	793	6526
P	0	2	47	6	5	4	3	67
S	183	292	5695	530	410	419	671	8200
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 31. Casos de Seguridad Social datos Confirmados.

SS	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
C	28	78	874	34	10	58	45	1127
E	0	0	1	0	0	1	0	2
N	46	140	1550	46	22	93	143	2040
P	0	1	20	0	0	1	0	22
S	60	187	2047	75	53	93	120	2635
Margen activo	134	406	4492	155	85	246	308	5826

ETNIAS QUE PRESENTAN CASOS

Otra parte de la información con que cuenta el set de datos es la etnia de la persona que fue infectada por el virus, en las siguientes ilustraciones 33, 34 y 35 se puede ver el comportamiento de los infectados con respecto a la etnia a que pertenece.

Ilustración 32. Etnias que reportan casos por año.

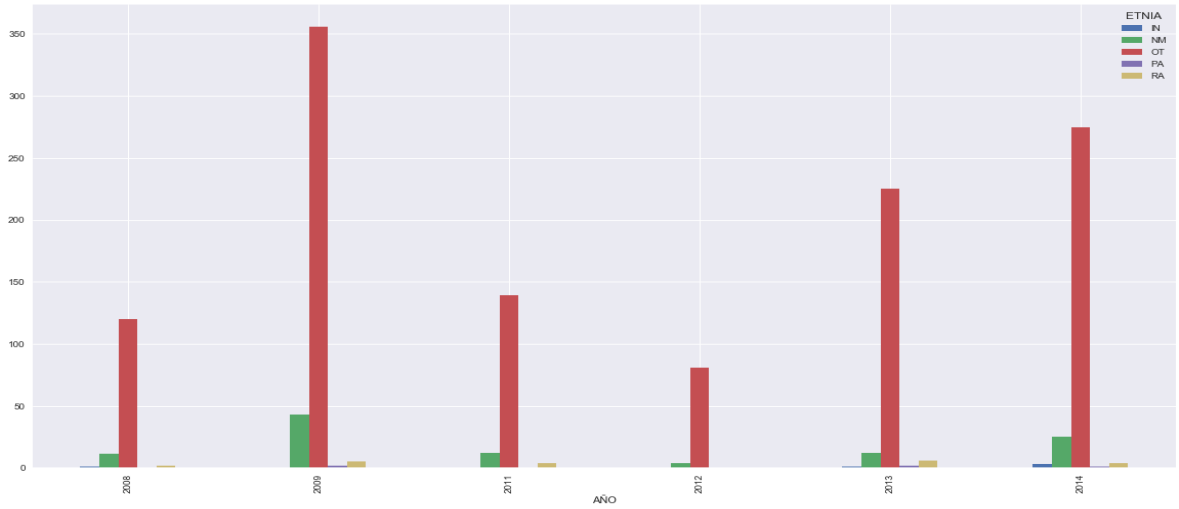


Ilustración 33. Casos anuales por Etnia casos Completos.

ETNIA	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
IN	3	0	49	4	0	7	10	73
NM	32	66	1104	76	57	104	154	1593
OT	367	582	11232	902	795	1030	1529	16437
PA	2	4	56	1	2	2	7	74
RA	4	8	221	15	10	18	33	309
RG	0	0	18	0	0	1	2	21
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 34. Casos anuales por Etnia casos Confirmados.

ETNIA	AÑOS							
	2008	2009	2010	2011	2012	2013	2014	Margen activo
IN	1	0	20	0	0	1	3	25
NM	11	43	388	12	4	12	25	495
OT	120	356	3998	139	81	225	275	5194
PA	0	2	21	0	0	2	1	26
RA	2	5	59	4	0	6	4	80
RG	0	0	6	0	0	0	0	6
Margen activo	134	406	4492	155	85	246	308	5826

OCUPACIÓN DE LAS PERSONAS INFECTADAS

El día a día que vive una persona con sus rutinas diarias tienden a mostrar el comportamiento de una persona y la ocupación de cada quien es un factor importante dado que puede determinar el lugar, el horario y las personas que suele frecuentar. En las siguientes ilustraciones 35 y 36 se puede ver el factor de coincidencia con las ocupaciones de los infectados:

Ilustración 35. Casos ocupaciones datos Completos.

	AÑOS							
OCUPACION	2008	2009	2010	2011	2012	2013	2014	Margen activo
AG	13	11	134	51	48	76	56	389
AR	20	21	282	26	21	30	60	460
EO	8	4	233	13	8	20	29	315
FP	0	0	12	5	1	9	23	50
OP	13	13	190	19	14	13	41	303
PE	1	2	48	6	6	6	13	82
PU	5	16	185	10	10	21	35	282
SV	3	7	261	23	16	26	58	394
TA	5	7	240	22	24	11	44	353
TN	340	579	11095	823	716	950	1376	15879
Margen activo	408	660	12680	998	864	1162	1735	18507

Ilustración 36. Casos ocupaciones datos Confirmados.

	AÑOS							
OCUPACION	2008	2009	2010	2011	2012	2013	2014	Margen activo
AG	6	6	39	8	3	22	8	92
AR	4	15	107	4	5	0	10	145
EO	3	2	86	1	2	7	4	105
FP	0	0	5	0	0	3	6	14
OP	4	9	58	0	1	2	6	80
PE	1	1	22	1	0	2	1	28
PU	2	10	56	2	0	4	7	81
SV	1	4	100	3	1	8	7	124
TA	0	5	88	3	1	2	7	106
TN	113	354	3931	133	72	196	252	5051
Margen activo	134	406	4492	155	85	246	308	5826

VII. EVALUACIÓN PRELIMINAR DE LAS CARACTERÍSTICAS DE LA INFORMACIÓN

Como primera instancia de las características de la información se encuentran algunas particularidades como se mencionan a continuación. Para el caso de las ilustraciones que contienen la información del tipo de género, la cantidad de casos para cada género es muy parejo, tanto así que el resultado equivalente final es del 51% de hombres infectados, por el 49% de mujeres infectadas, lo cual no es muy diferencial desde el punto de vista estadístico descriptivo, por lo que se podría afirmar como hipótesis que la infección no tiene preferencia de género y por eso el sexo de la persona no es un factor determinante para la transmisión del virus dengue con el análisis preliminar mas no con un análisis estadístico profundo. Para la ilustración del género donde los casos son confirmados la tendencia sigue siendo la misma dado que la proporción es 50.5% hombres infectados contra 49.5% de mujeres infectadas.

Las ilustraciones de localización de las zonas, muestra donde el vector se establece con más facilidad y es realmente resaltable dado que los casos se presentan en 75.5% de proporción en las cabeceras municipales por lo cual puede ser uno de los aspectos a tener en cuenta al momento de tratar de erradicar o prevenir el dengue.

La información concerniente al grupo etario muestra al parecer que la edad tiene cierta relación con la tasa de infectados, se muestra en las ilustraciones correspondientes que para cada año los grupos etarios A y B representan el 60% o superior entonces se entiende que estos tienden a sufrir por el virus en mayor proporción y que en cada año se mantiene esta tendencia.

Al observar las ilustraciones sobre el estrato social, es reconocible que la mayoría de los casos se presentan en los estratos D, E, y F con un porcentaje superior al 90% en toda la línea de tiempo, y en especial el más afectado es el estrato F pues en casi todos los años aporta un valor cercano al 50% de la totalidad de la muestra, entonces es posible que el vector se sienta mucho más cómodo en zonas de bajos recursos económicos.

En las ilustraciones sobre la seguridad social se percibe otra característica destacable y es que existe un alto nivel de coincidencia del virus con las personas que no cuentan con el sistema de salud o que pertenecen al régimen de salud subsidiado llegando casi en todos los casos por año con un 80% o superior, esta población es la que más reporta casos de dengue, aunque las personas que pertenecen a los demás regímenes también aportaron casos, pero en una cifra bastante lejana comparada con el resto de los casos.

Las ilustraciones sobre ocupaciones muestran que se tienen por parte de los sujetos con el virus una tendencia fuertemente sesgada hacia el lado de las personas que tienen ocupación TN reportando como mínimo el 80% de los casos por año, mostrando que esta ocupación es muy interesante para que el vector se establezca, reaparezca o se mantenga. Y por último la información de las ilustraciones referentes a las etnias preferentes en los casos de dengue, en la cual por excelencia el mosquito contagia a los tipos OTROS con una incidencia del 90% anual.

Teniendo como referencia todo lo evaluado anteriormente es un buen indicio las propiedades importantes en el set de datos, que ayudan a inclinar la decisión de cuáles

algoritmos pueden ser beneficiosos al usar el análisis con machine learning y así obtener mejor provecho de los datos.

Para efectos del análisis se hará la evaluación sobre el set de datos con casos confirmados y no confirmados, logrando observar su comparación, y como en el año 2010 hubo un brote anormal en relación a los demás años y está registrado en un boletín titulado “Alerta Epidemiológica: Actualización sobre brote de Dengue en las Américas”, publicado y actualizado periódicamente en 2010 por la OMS y la OPS (OPS, Organización Panamericana de la Salud, 2010), es conveniente trabajar independientemente por años para evitar posibles inclinaciones sobre un punto específico que puede ser generado por los datos de los años con mayor número de casos reportados del virus en cuanto a porcentaje se refiere.

5. ANÁLISIS DE LA INFORMACIÓN

Para analizar toda la información que se tiene se debe elegir la opción más adecuada para desarrollar esta tarea, entonces se someterán a evaluación los algoritmos que fueron nombrados en el estado del arte, con los cuales se hicieron estudios y arrojaron resultados bastante considerables en la mayoría de los casos y que se mencionarán en la siguiente sección.

I. ANÁLISIS RELACIONAL DEL SET DE DATOS

Una de las maneras que ayuda a identificar información relevante en un conjunto de datos, es conocer si existe algún tipo de relación entre las variables que lo conforman, para esto se usa una técnica como por ejemplo, el análisis de correspondencia simple o también el análisis de correspondencia múltiple, el cual permite hallar la correlación de unas variables con respecto a otras. Para este set de datos que se tiene es muy conveniente, dado que el análisis de correspondencia se basa en la relación de variables categóricas y a diferencia de la edad, el año, el mes y la semana que son variables numéricas y que la edad posteriormente fue transformada en una variable categórica llamada grupo etario, se podría afirmar que el set de datos está dominado por el tipo categórico, lo que es un buen aporte para obtener un análisis viable de correlación de la muestra, entonces como variable de agrupación para el estudio, se tomará como base de estudio el municipio, ya que este nos permite dar un vistazo de manera local por las zonas donde se transporta el vector.

La forma en que se determina si existe o no algún tipo de relación entre variables, se logra contrastando hipótesis basado en la independencia de las variables en cuestión. Regularmente el test utilizado para esto, es el del Chi-cuadrado de Pearson, Se evalúa la hipótesis nula que asume anticipadamente la independencia entre ambas variables, mediante el estadístico X^2 de Pearson. Donde H_0 o hipótesis nula se toma que ambas variables son independientes, y H_1 propone que existe relación de dependencia. La prueba compara los perfiles de las filas y columnas con los perfiles marginales correspondientes, teniendo en cuenta que si H_0 es verdadera la totalidad de los perfiles fila (respecto columna) son similares entre sí e iguales al perfil marginal de las filas con respecto a las columnas. Cuando la significancia es menor a 0.05 se rechaza la hipótesis nula y se concluye que existe una fuerte relación de dependencia entre las variables, en el caso de que la significancia supere el 0.05 se acepta la hipótesis nula y se considera que las variables son independientes entre ellas. Las dimensiones de una solución permiten explicar la mayor parte de la variación, la cantidad máxima de factores que existen para las dimensiones están determinados por el número total de características distintas de la variable menos 1. Por ejemplo, la variable género tiene dos valores o factores diferentes que son: masculino y femenino, entonces el máximo número de dimensiones para este caso es 2 características menos 1, para un total de una dimensión. La significancia en las tablas se verá representado por la columna ‘sig’. A continuación, se exponen los análisis de cada variable respecto al municipio.

La siguiente ilustración 37 muestra el análisis de correspondencia para la variable género con respecto al municipio.

Ilustración 37. Análisis de Correspondencia Municipio vs Género.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza
					Contabilizado para	Acumulado	Desviación estándar
1	,046	,002			1,000	1,000	,013
Total		,002	12,570	,561[a]	1,000	1,000	

a. 14 grados de libertad

El número de dimensiones máximo para la comparación de la variable género es 1, como se había indicado con anterioridad, esa sola dimensión es suficiente para explicar la variación de los factores en esta dimensión y la significancia es 0.561 con lo que se afirma la hipótesis nula, por lo cual se concluye que no existe relación de dependencia entre la variable género y la variable municipio, que de alguna manera confirma la hipótesis propuesta en la evaluación preliminar que tomaba como referencia que la muestra de casos de dengue está distribuida en cantidades similares tanto para hombres como para mujeres.

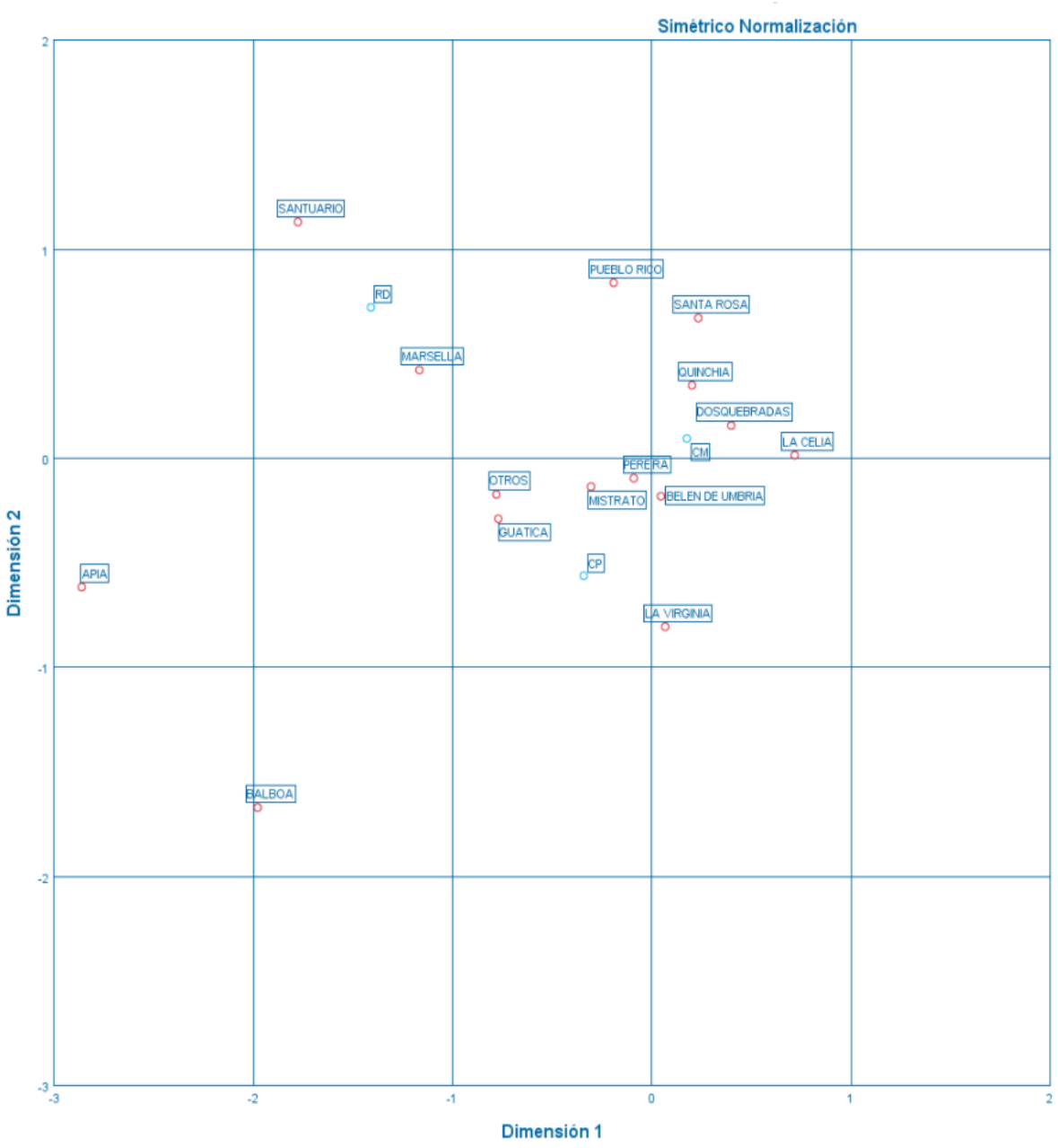
Ilustración 38. Análisis de Correspondencia Municipio vs Tipos de Área.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,142	,020			,703	,703	,017	,194
2	,092	,009			,297	1,000	,014	
Total		,029	167,386	,000[a]	1,000	1,000		

a. 28 grados de libertad

Para las áreas se cuenta con dos dimensiones de análisis, en la cual la significancia es de 0, permitiendo concluir que los casos de dengue por ciudad están relacionados directamente por las zonas donde ocurren los casos y era de preverse dado que en las tablas descriptivas de esta variable se observaba que existen áreas específicas en cada ciudad, con mayor tendencia a presentar los casos.

Ilustración 39. Análisis dimensional Área vs Municipio.



En la ilustración 39, por dimensiones, se muestra la dispersión de las áreas donde habitan las personas, contra los municipios donde la cabecera municipal CM es la que tiende a mantener mayor proporción de los municipios de la muestra.

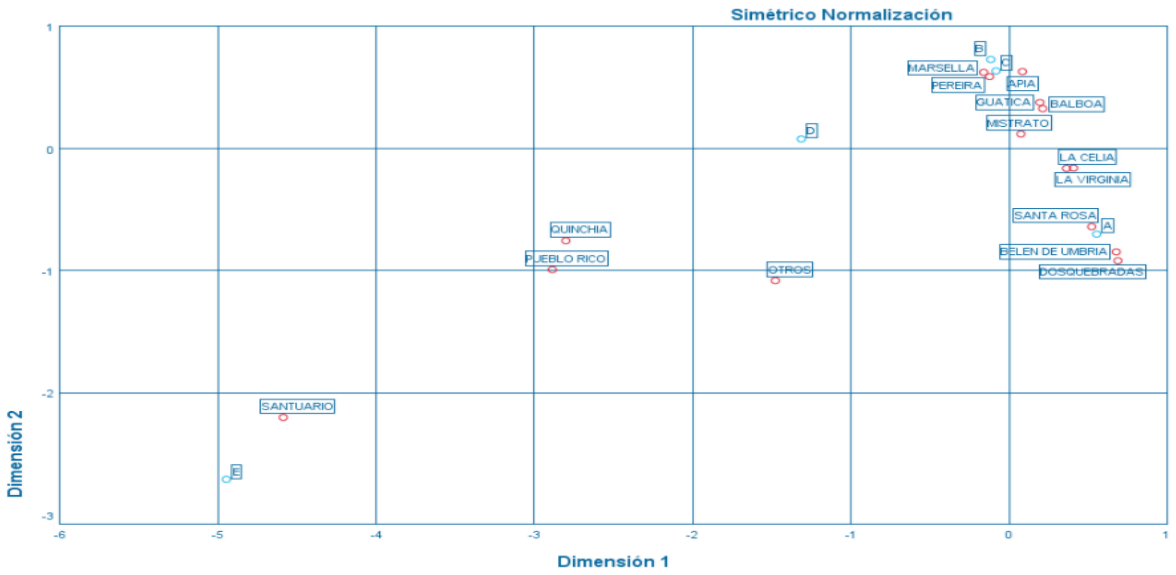
Para los grupos etarios en la ilustración 40 se puede observar que las primeras dos dimensiones tienen la capacidad de explicar el 96.4% de la variabilidad del modelo, que también se puede apreciar con los valores de la inercia y que dado el valor de la significancia se puede concluir que existe relación dependiente entre el municipio y el grupo etario de las personas que son afectadas por el virus.

Ilustración 40. Análisis de Correspondencia Municipio vs Grupo Etario.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,655	,429			,562	,562	,015	,575
2	,553	,306			,402	,964	,011	
3	,125	,016			,020	,984		
4	,109	,012			,016	1,000		
Total		,762	4440,290	,000[a]	1,000	1,000		

a. 70 grados de libertad

Ilustración 41. Análisis dimensional Municipio vs Grupo Etario.



La parte dimensional de la ilustración41 permite entender que los patrones de afectación son distintos según el municipio y el grupo etario donde en gran cantidad de municipios los casos están entre los grupos etarios A, B y C, en municipios como Santuario está inclinado al tipo E.

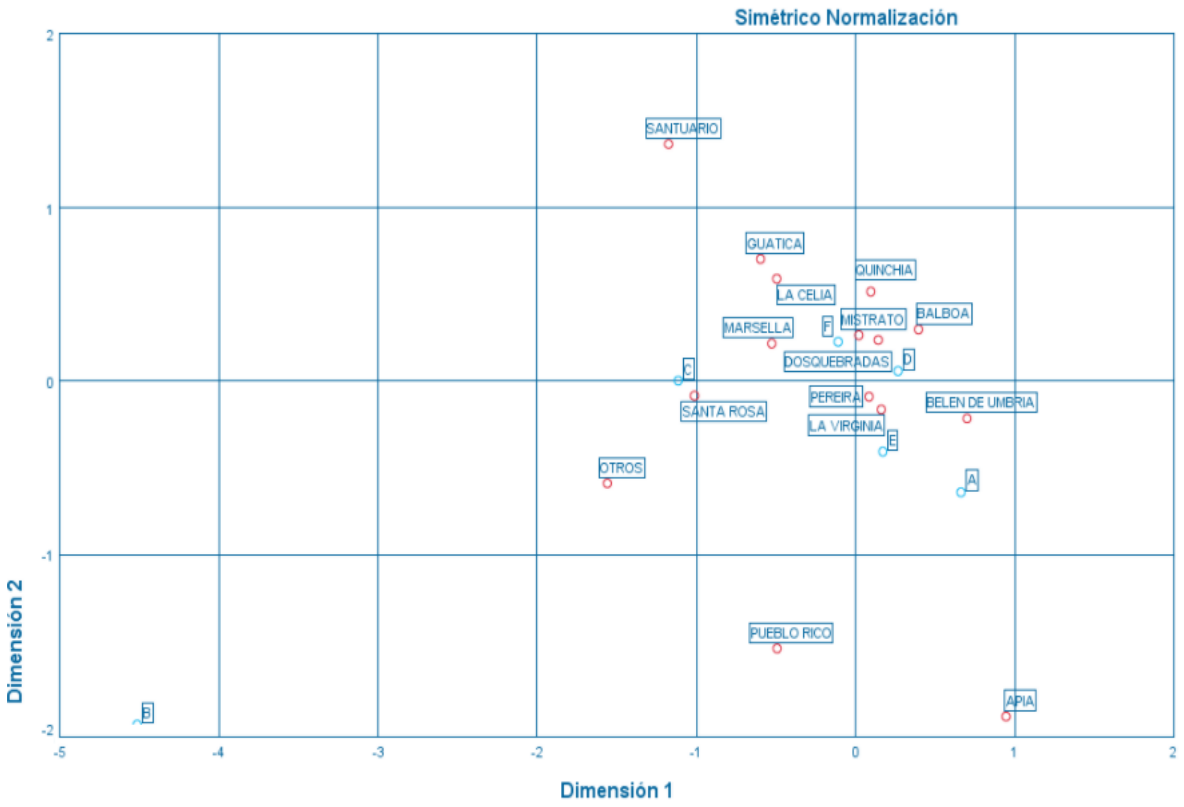
Los estratos sociales necesitan de 4 dimensiones para explicar el 97.2% de la variabilidad en la muestra, era algo de esperarse basado en los datos obtenidos con anticipación, la significancia indica que existe relación directa entre ambas variables.

Ilustración 42. Análisis de Correspondencia Municipio vs Estratos.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,137	,019			,553	,553	,021	,038
2	,085	,007			,211	,764	,014	
3	,064	,004			,121	,885		
4	,054	,003			,087	,972		
5	,031	,001			,028	1,000		
Total		,034	197,639	,000[a]	1,000	1,000		

a. 70 grados de libertad

Ilustración 43. Análisis dimensional Municipio vs Estratos.



La ilustración 43 dimensional muestra que en gran parte de los municipios están muy cercanos los casos para los estratos C, D, E y F.

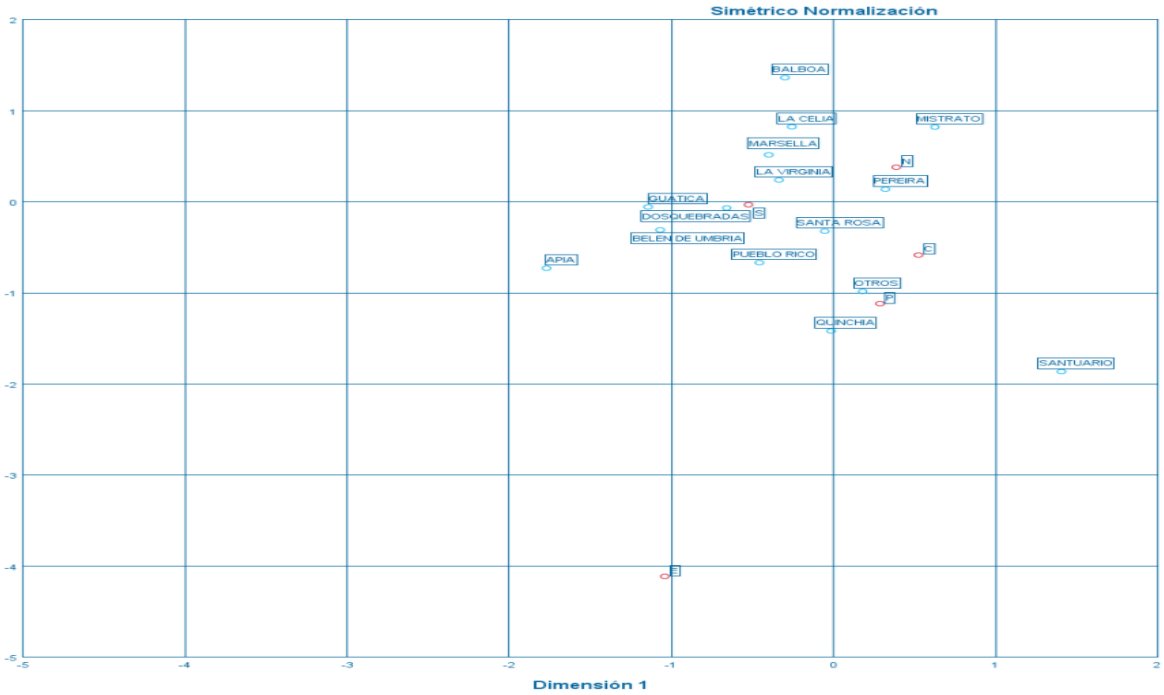
La seguridad social muestra que las primeras dos dimensiones son bastante significativas para explicar la variabilidad del modelo, pero la inclusión de la tercera dimensión permite explicar casi toda la variabilidad del modelo. Y la significancia indica relación directa entre las variables.

Ilustración 44. Análisis de Correspondencia Municipio vs Seguridad Social.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,231	,053			,699	,699	,012	,069
2	,128	,016			,214	,913	,015	
3	,061	,004			,049	,962		
4	,054	,003			,038	1,000		
Total		,076	443,110	,000[a]	1,000	1,000		

a. 56 grados de libertad

Ilustración 45. Análisis dimensional Municipio vs Seguridad Social.



La dimensionalidad de la ilustración 45, muestra la variabilidad de los casos por zonas donde se reportaron los casos, explica que algunos municipios cuentan con su particularidad de seguridad social para los infectados.

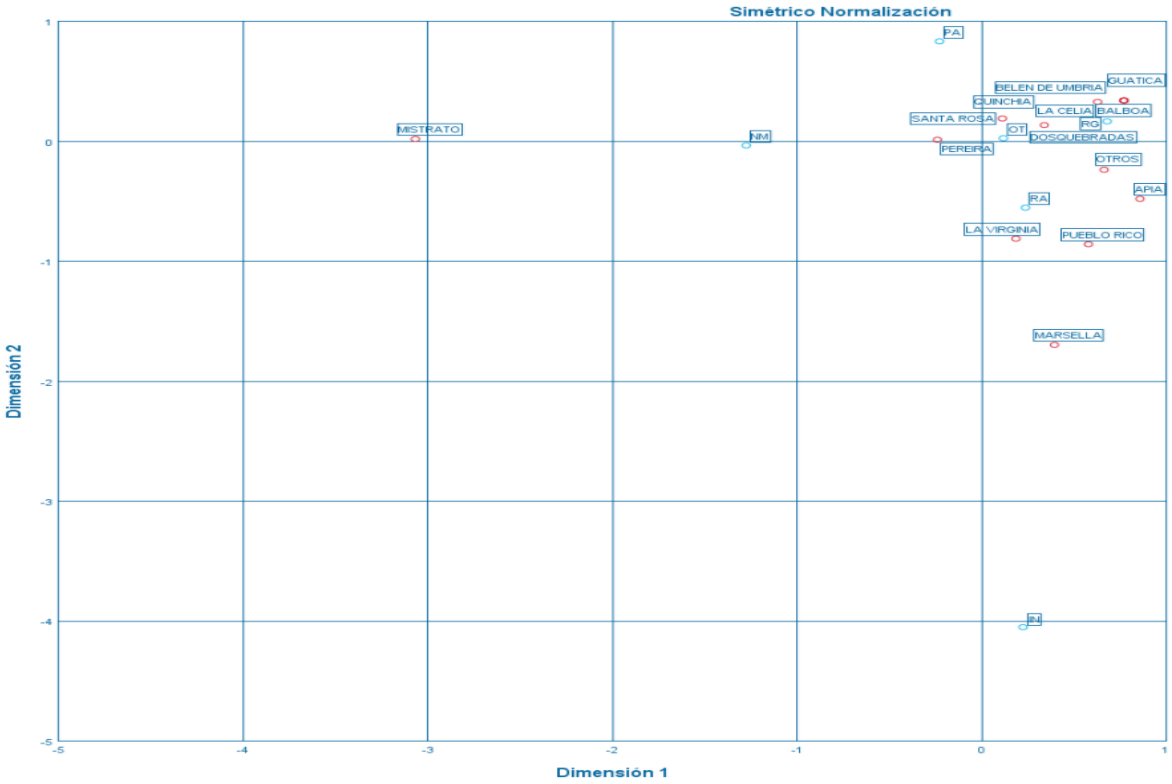
La ilustración 46 muestra que la significancia indica que existe una relación entre las etnias y los municipios y que tres dimensiones pueden presentar un alto nivel de comprensión del 95.5% en la variabilidad de la muestra.

Ilustración 46. Análisis de Correspondencia Municipio vs Etnias.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,152	,023			,678	,678	,016	,012
2	,078	,006			,180	,858	,025	
3	,057	,003			,096	,955		
4	,036	,001			,037	,992		
5	,017	,000			,008	1,000		
Total		,034	199,068	,000[a]	1,000	1,000		

a. 70 grados de libertad

Ilustración 47. Análisis dimensional Municipio vs Etnias.



Las dimensiones de la ilustración 47 muestran que las etnias están cuentan con mayor centralización para la mayoría de los municipios, en especial la etnia OT.

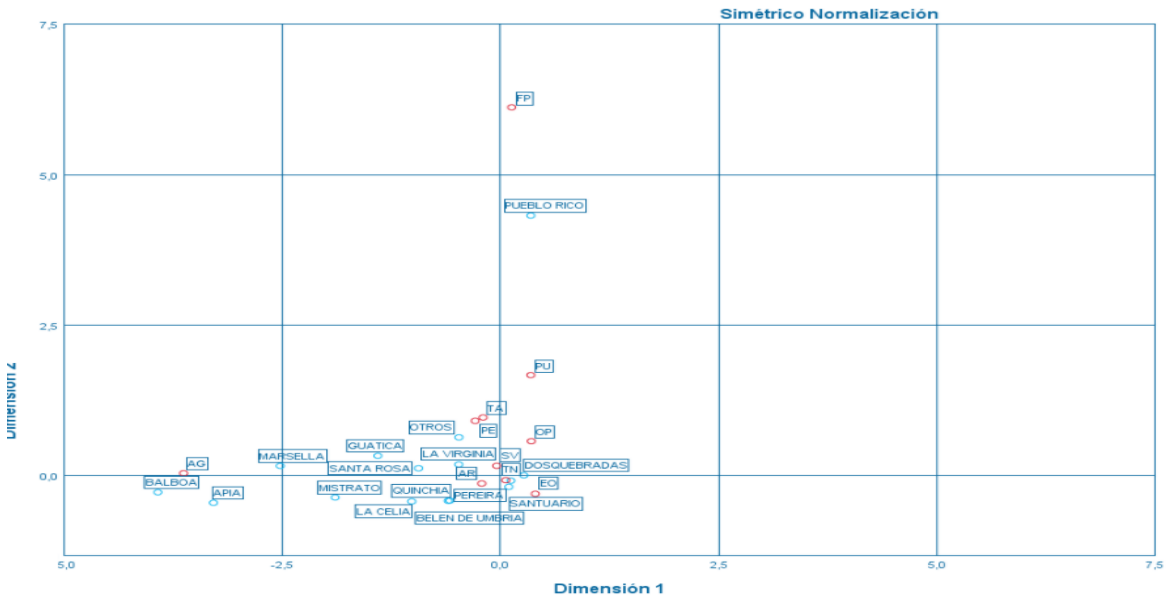
Las ocupaciones de las personas afectadas con el virus, ilustración 48 usando tres de las nueve dimensiones pueden explicar un 90.8% de la variabilidad de la muestra y la significancia indica una relación entre las variables municipio y ocupación.

Ilustración 48. Análisis de Correspondencia Municipio vs Ocupación.

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,220	,048			,542	,542	,028	,001
2	,162	,026			,294	,835	,045	
3	,081	,007			,073	,908		
4	,048	,002			,026	,934		
5	,046	,002			,023	,958		
6	,043	,002			,020	,978		
7	,034	,001			,013	,991		
8	,027	,001			,008	,999		
9	,009	,000			,001	1,000		
Total		,089	519,738	,000[a]	1,000	1,000		

a. 126 grados de libertad

Ilustración 49. Análisis dimensional Municipio vs Ocupación.



Los distintos patrones en las dimensiones de la ilustración 49 para cada región, indican que los oficios relacionados en cada municipio con el caso del dengue tienden a ser muy similares en algunos municipios, pero en otros municipios tienen otras ocupaciones que pueden ser afectadas con mayor frecuencia.

II. ELECCIÓN DE CLASE PARA EL ANÁLISIS DE DATOS CON MACHINE LEARNING

Como el machine learning se basa en aprender de los datos y luego predecir qué podría suceder con datos hipotéticos o reales para su verificación, es fundamental seleccionar una característica que sea clave para el análisis de datos, en este caso se tendrá en cuenta como base la información de los datos, entre estos destaca con una gran ventaja la variable tiempo, debido a que existe información anual, mensual y semanal, y como anteriormente se mencionó debido a que en el 2010 hubo una multiplicación de casos del virus del dengue de manera inusual, se decidió fraccionar la información anualmente, entonces queda la característica, mensual y semanal, es posible hacer una serie de tiempo con los datos que se tienen y regularmente entre mayor detallada se tenga la información mejor resultados podría arrojar el modelo y una posible comprensión superior del fenómeno a estudiar, por eso se tendrá en cuenta como la clase principal MES dado que puede aportar buena sensibilidad al modelo, pues cuenta el año con 12 meses y las semanas en la mayoría de los casos son 53 para un periodo anual, existen varios casos en el set de datos donde no existe reporte por varias semanas y eso altera la fiabilidad del modelo. Entonces todos los análisis posteriores tendrán como base principal la variable tiempo, representada por la clase mes.

III. SELECCIÓN DEL MÉTODO DE VALIDACIÓN EN ALGORITMOS DE MACHINE LEARNING

Una de las características fundamentales para el éxito de un análisis de predicción con machine learning es la selección del método de validación adecuado para el conjunto de

datos que se tiene, y es cuando el modelo está listo para ser evaluado, normalmente sigue la etapa de entrenamiento del conjunto de datos, y luego se verifica si el modelo tuvo un nivel de aprendizaje aceptable para la muestra. Este tema de verificación se realiza mediante un método de validación, regularmente se usan los métodos de validación más populares para esta tarea, uno es el método de la validación cruzada o cross validation y el otro es el método de retención o holdout method. La validación cruzada propone fraccionar el set de datos en el número de segmentos que se quiera, en la cual cada parte contiene la misma cantidad de registros. El método de retención divide la muestra básicamente en dos partes, una es el set de datos de entrenamiento y la otra es el set de datos de validación de la información. Ambas técnicas son de alto rendimiento y arrojan resultados muy buenos cuando se hace la elección correcta para el set de datos que se quiere analizar.

En este caso particular la validación cruzada tiene una leve restricción que se basa en que, al momento de dividir la muestra, esta no debe superar la cantidad de miembros totales que contiene una clase, los miembros se consideran valores distintos dentro de la misma clase, aquí sale la primera restricción de fraccionamiento de la información dado que dentro del set de datos tanto completo como de datos confirmados existe una clase llamada género que contiene dos miembros dentro de ésta que corresponden a los valores masculino y femenino, por esta razón el set de datos solo es posible dividirlo en dos conjuntos que en el mayor de los casos no se recomienda si se busca tener un buen nivel de clasificación. Si se quiere aumentar la cantidad de divisiones de la información es necesario prescindir de la variable género y usar por ejemplo el área que es la que

sigue en cantidad de valores mínimos para este caso tres valores CM, CP, RD, pero es posible que la variable género aporte una buena cantidad de porcentaje de éxito de clasificación al algoritmo que se esté usando, o alguna de las variables que se prescindan pueden ser importantes para el set de datos. Si se usa el método de retención, el validador fracciona los datos que existen y trabaja con ellos sin preocuparse por la cantidad de miembros que contengan cada una de las clases, ya que éste usa una parte del set de datos para entrenar la máquina y aprender de los datos, y la cantidad de datos restantes se utiliza para verificar qué tan bien aprendió el algoritmo de los datos que usó como entrenamiento, entonces para no prescindir de ninguna clase del set de datos, se usará como método de validación predeterminado el método de retención, este será configurado con un porcentaje de entrenamiento del ochenta por ciento del total del set de datos y el otro veinte por ciento de los datos para la validación, tanto para los datos de casos completos, así como para los datos del set de datos confirmados. Entonces todo el análisis de los datos con los algoritmos se realizará aplicando el método de retención con proporción ochenta – veinte.

IV. ALGORITMOS DE MACHINE LEARNING PARA EL SET DE DATOS

Al tener conocimiento de la información con que cuenta el set de datos a estudiar con machine learning, se hace necesario elegir un algoritmo favorable que permita estudiar el fenómeno con un nivel de aprendizaje alto y un buen porcentaje de acierto en la predicción del clasificador seleccionado, por este motivo se pondrán a estudio los algoritmos que se mencionaron en el estado del arte, los clasificadores utilizados en esos estudios son una buena referencia ya que han sido usado en estudios de temas

epidemiológicos y fenómenos similares. Los algoritmos que fueron elegidos son en total diez, se pondrán a prueba en su estado por defecto contra el set de datos que se tiene y se evaluarán contra el set de datos completo y los del set de datos confirmados, en total se entrenarán diez máquinas de aprendizaje y se evaluará el comportamiento del set de datos con cada uno de los algoritmos seleccionados, posteriormente se elegirá cuál de las máquinas de aprendizaje favorece en mayor proporción la explicación del fenómeno, los algoritmos de clasificación que fueron elegidos para la prueba en el conjunto de datos contienen el nombre del clasificador y su simbología, estos se muestran listados en la siguiente tabla 52.

Tabla 22. Algoritmos Seleccionados para evaluación del set de datos.

NÚMERO	ALGORITMO	SÍMBOLO
1	Support Vector Machine Regressor	SVMR
2	Support Vector Machine Classifier	SVMC
3	Logistic Regression	LR
4	Linear Model	LM
5	MLP Classifier	MLPC
6	MLP Regressor	MLPR
7	Gradient Boosting Classifier	GBC
8	Linear model Lasso	LASSO
9	Decision Tree Classifier	DT
10	Random Forest Classifier	RF

Seguidamente se realizará la evaluación del set de datos con todos los algoritmos mencionados en la tabla anterior, esto con el fin de comparar su precisión y así seleccionar con qué modelo se evaluará la muestra finalmente. Los gráficos siguientes de las ilustraciones 50 a la 61 muestran en valores porcentuales, que tanto es capaz de aprender cada algoritmo de la muestra que se tiene, en este caso de ambas muestras, la del set de datos completo y del set de datos confirmado.

Ilustración 50. Rendimiento algoritmo SVMC. Ambos sets de datos.

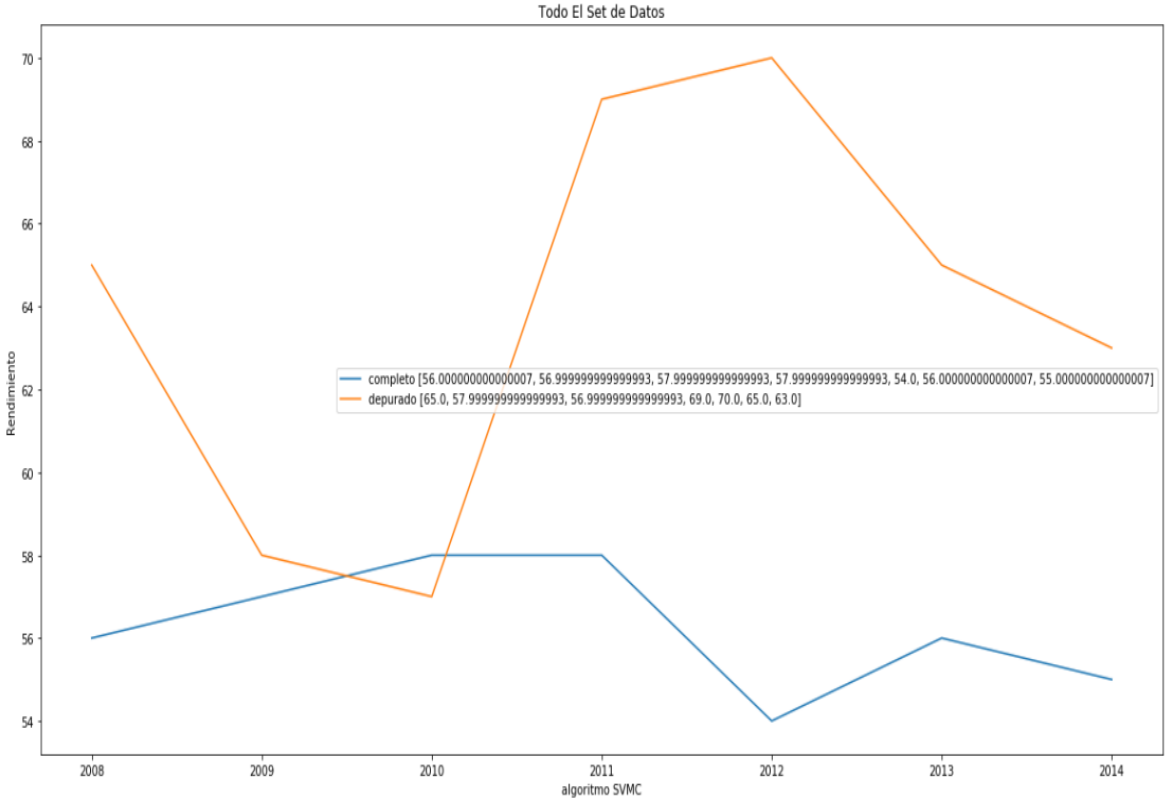


Ilustración 51. Rendimiento algoritmo SVMR. todos los datos.

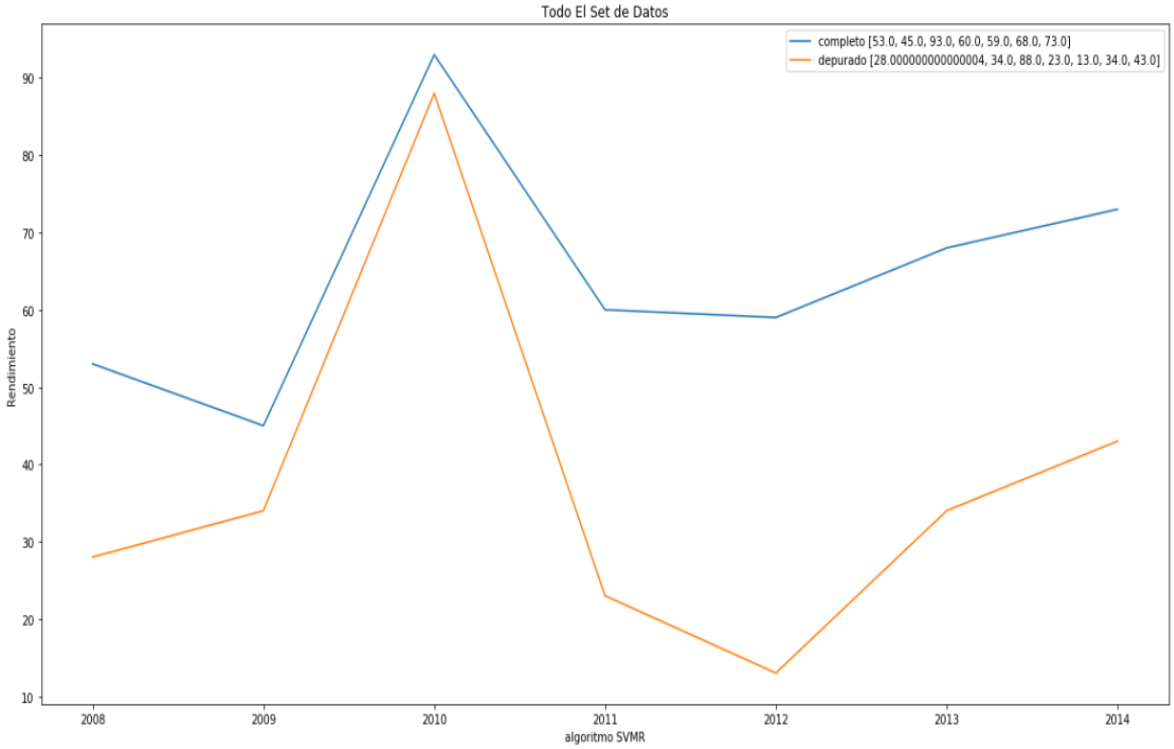


Ilustración 52. Rendimiento algoritmo LM. Ambos sets de datos.

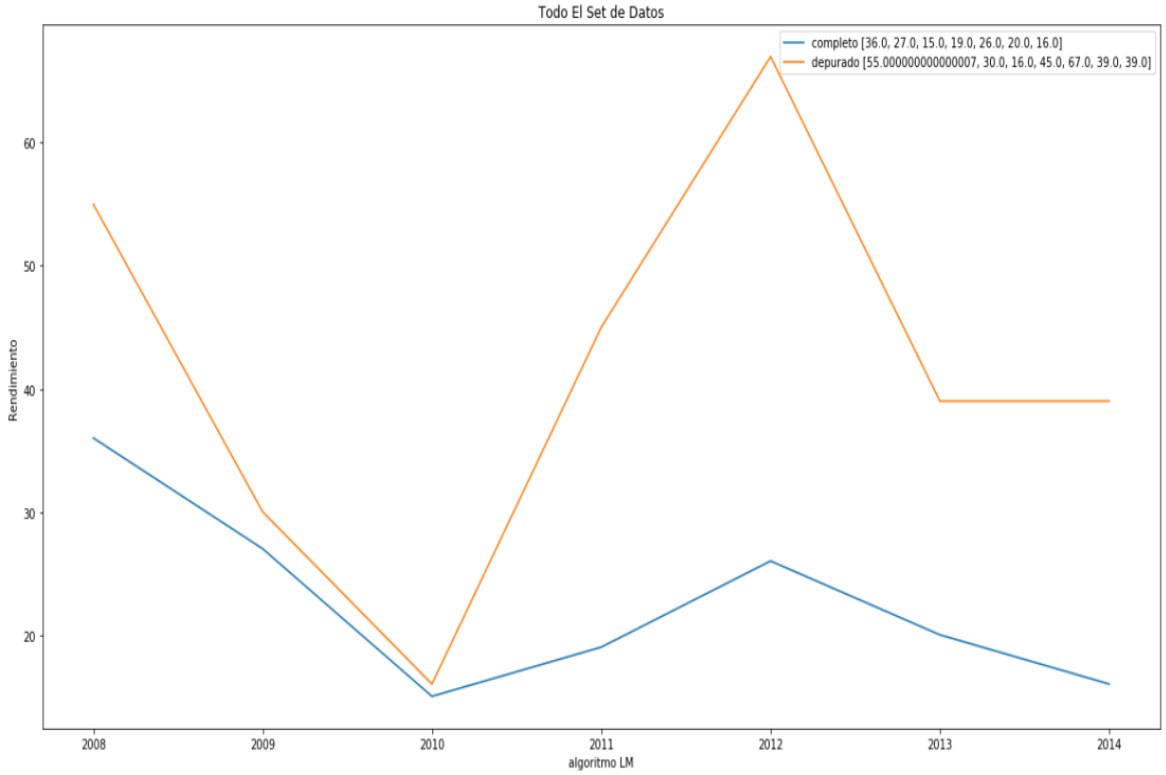


Ilustración 53. Rendimiento algoritmo LR. Ambos sets de datos.

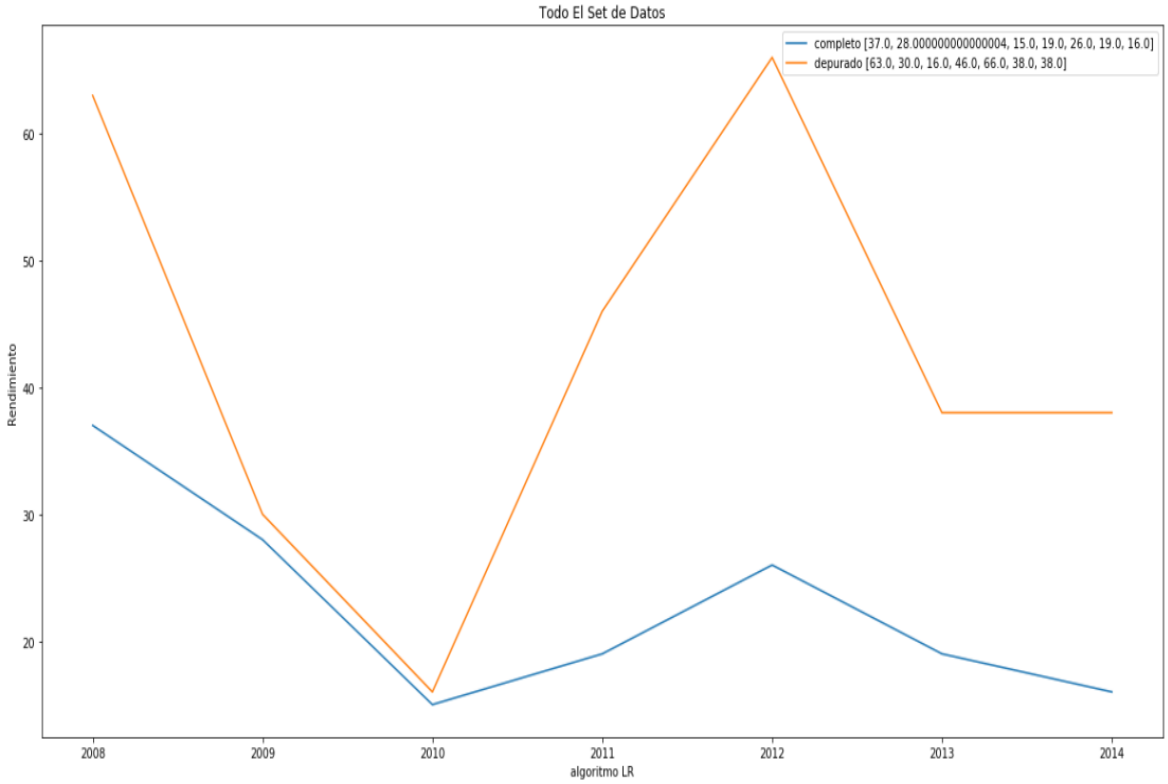


Ilustración 54. Rendimiento algoritmo MLPC. Ambos sets de datos.

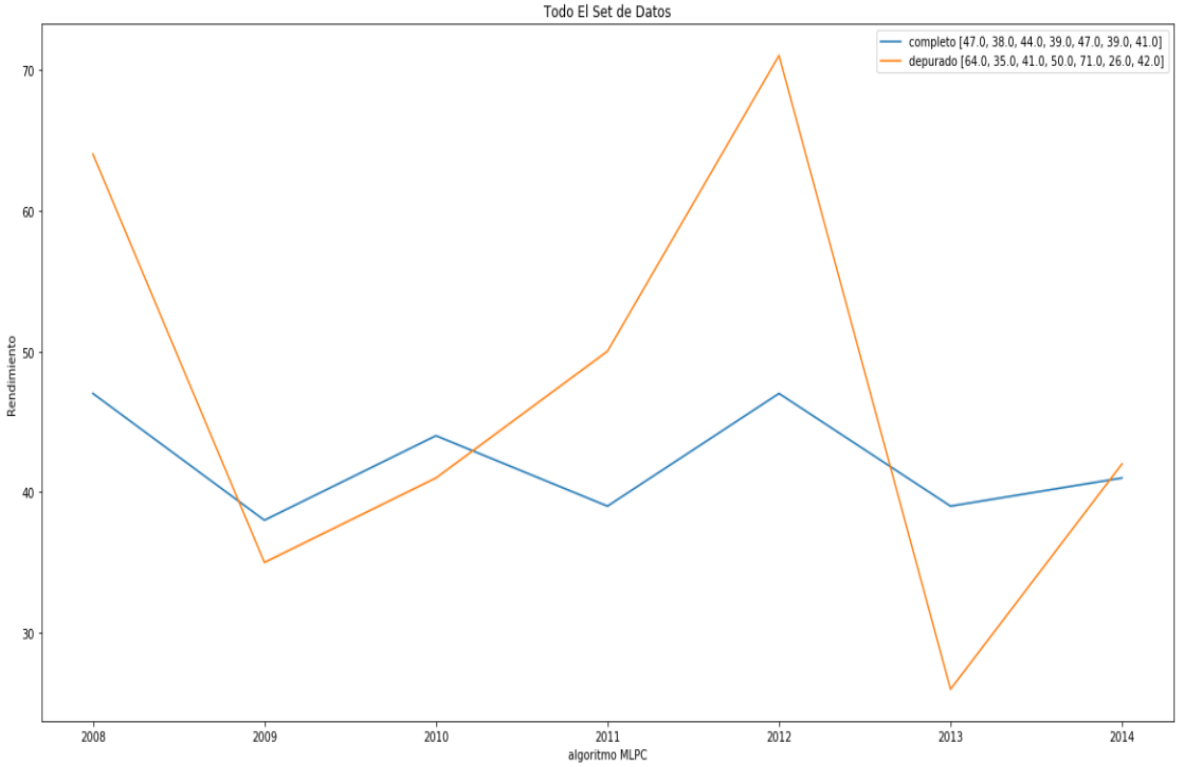


Ilustración 55. Rendimiento algoritmo MLPR. Ambos sets de datos.

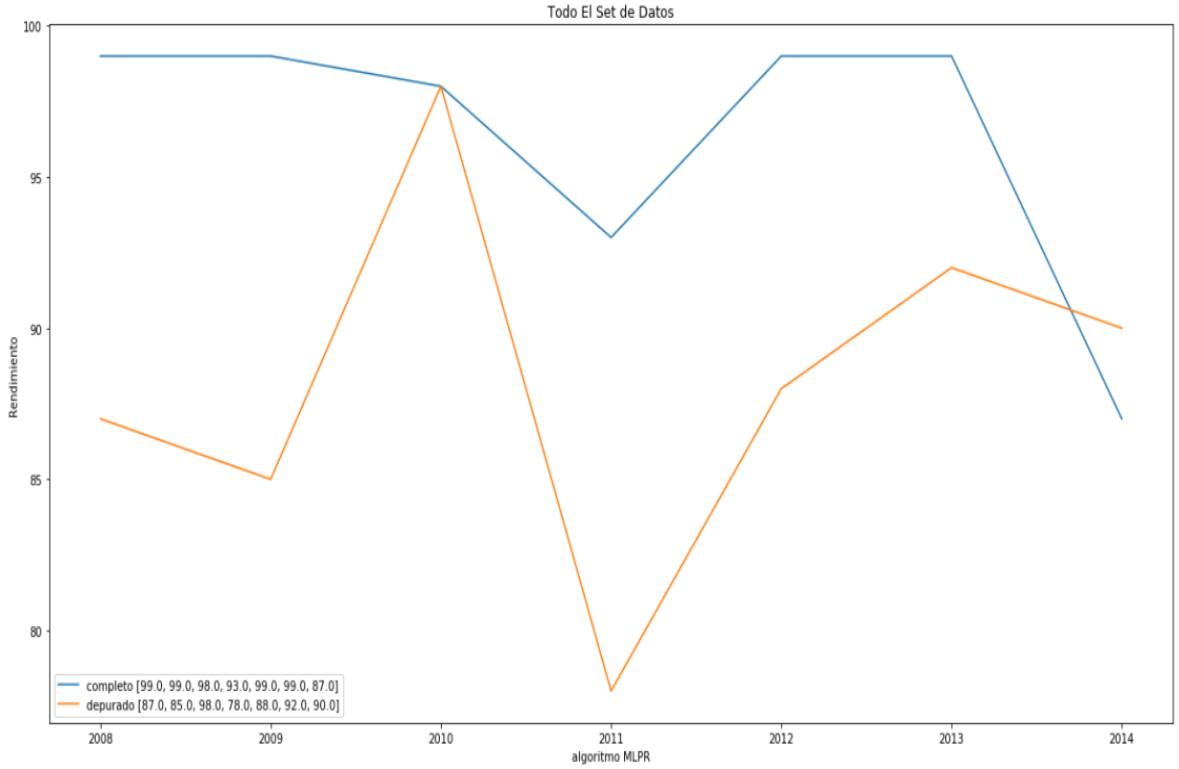


Ilustración 56. Rendimiento algoritmo GBC. Ambos sets de datos.

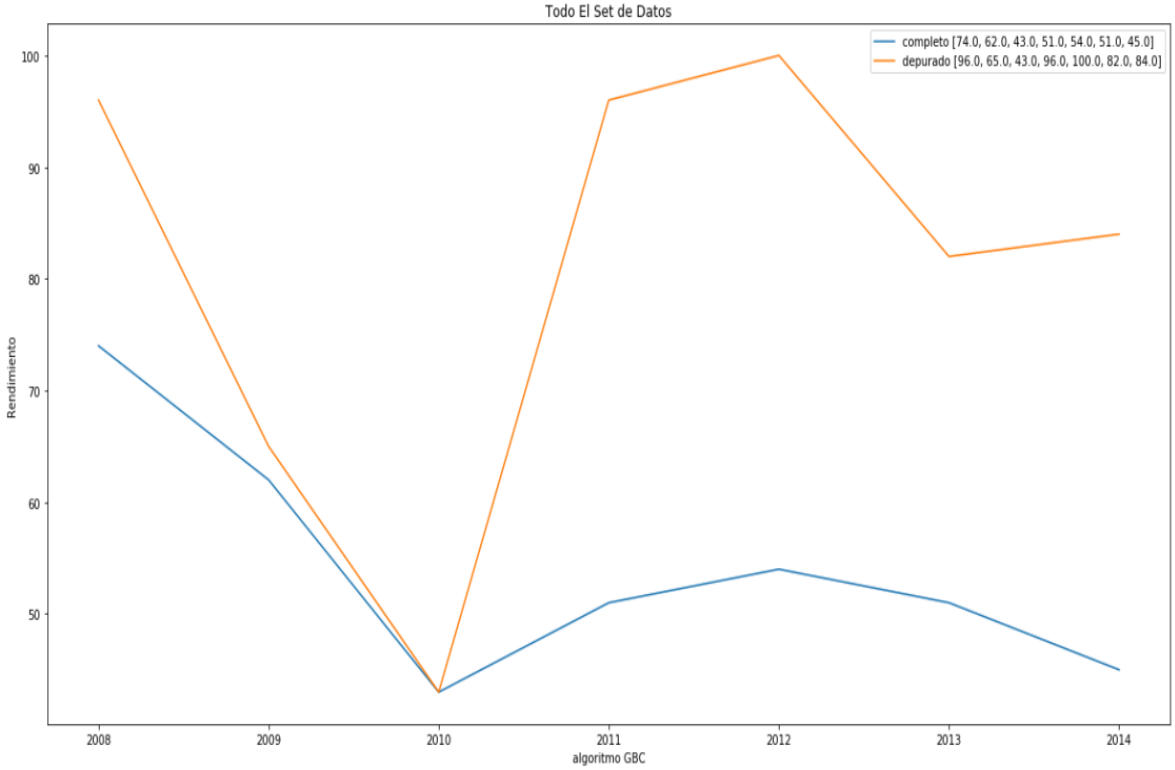


Ilustración 57. Rendimiento algoritmo LASSO. Ambos sets de datos.

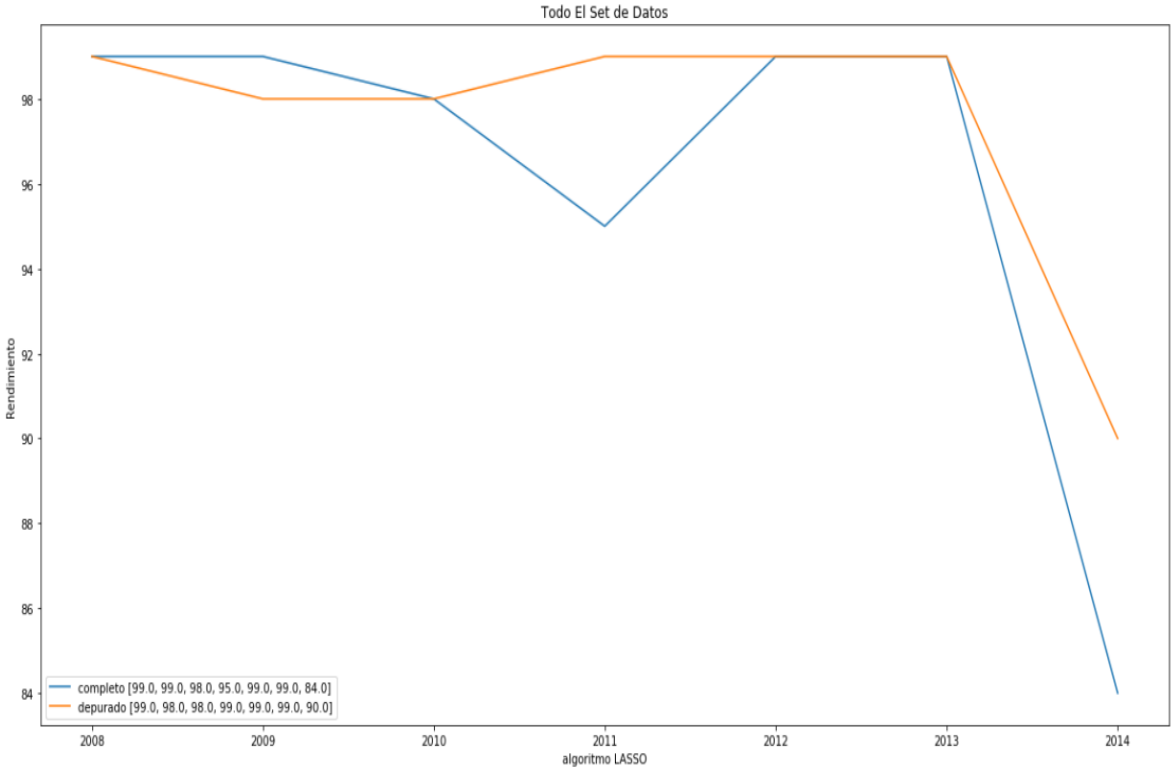


Ilustración 58. Rendimiento algoritmo DT. Ambos sets de datos.

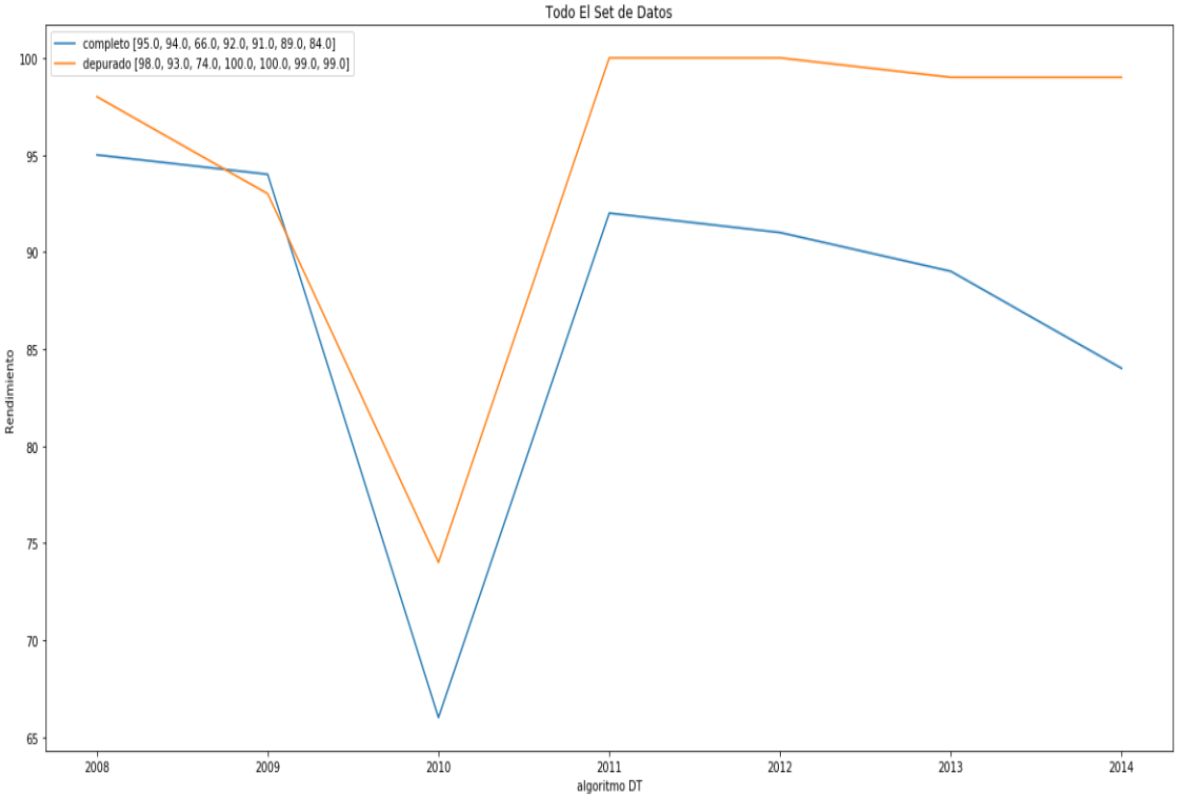


Ilustración 59. Rendimiento algoritmo RF. Ambos sets de datos.

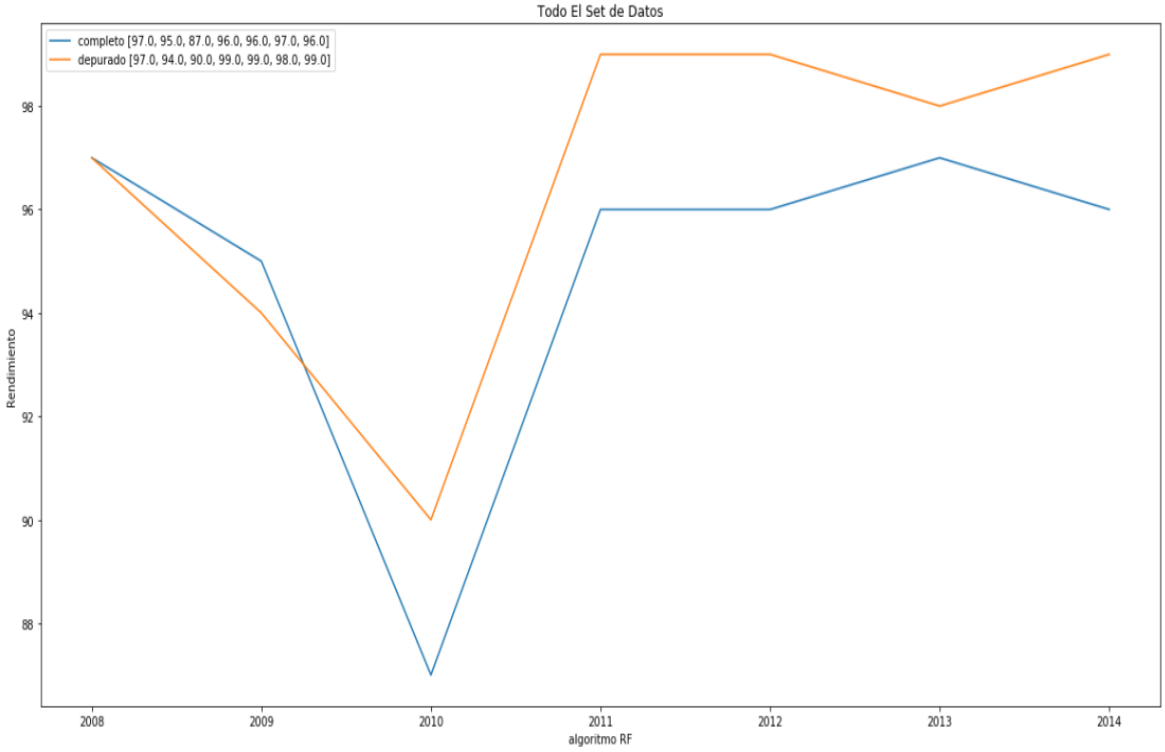


Ilustración 60. Comparación algoritmos set de datos completo.

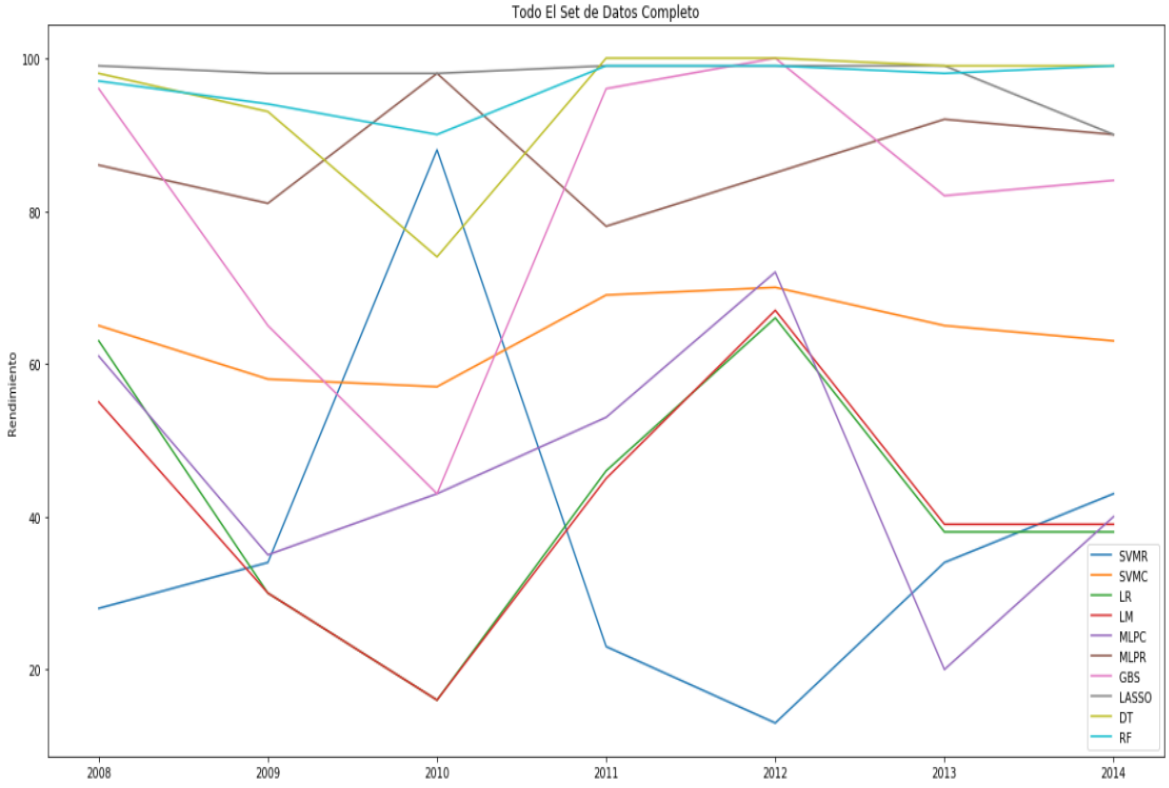
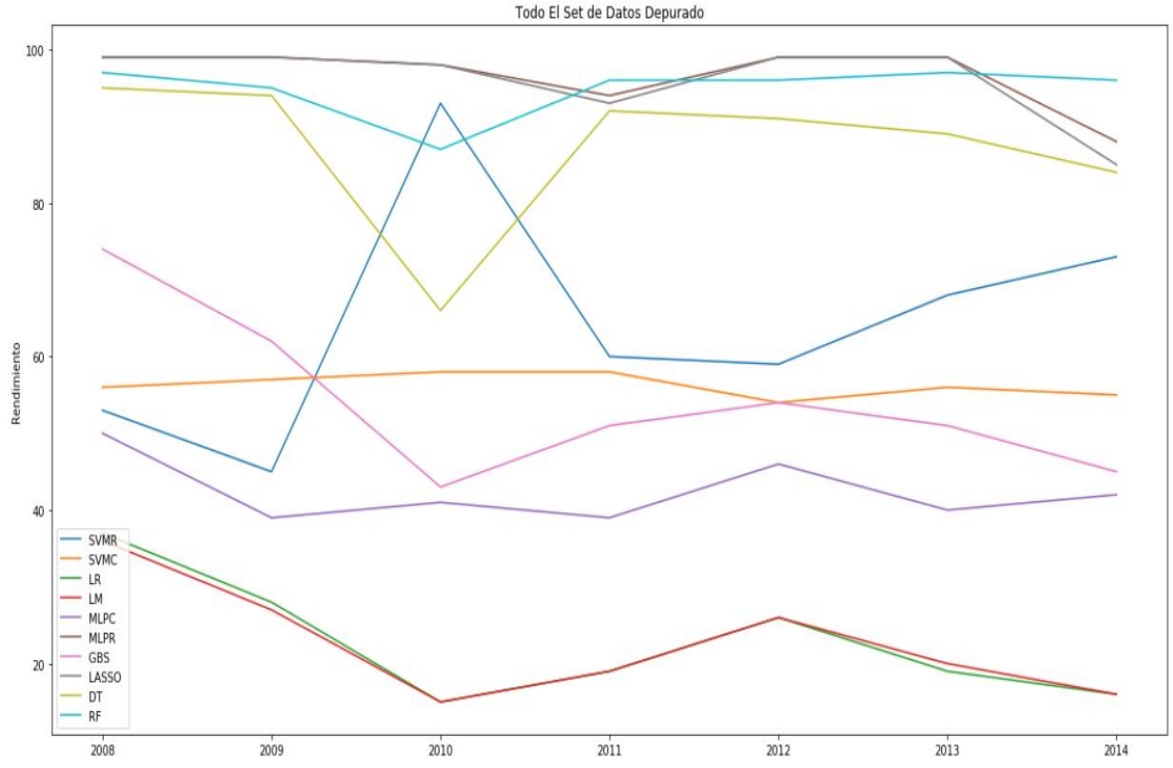


Ilustración 61. Comparación algoritmos set de datos confirmados.



V. EVALUACIÓN DE LOS ALGORITMOS DE APRENDIZAJE

En las tablas siguientes se hace un resumen anual, de todos los algoritmos que se utilizaron en cuenta para la evaluación del set de datos tanto completo, así como confirmado, tal y como se muestra a continuación en las tablas 23 y 24.

Tabla 23. Evaluación del rendimiento del estimador del algoritmo anual set de datos completo.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
SVMR	53	45	93	60	59	68	73
SVMC	56	57	58	58	54	56	55
LR	37	28	15	19	26	19	16
LM	36	27	15	19	26	20	16
MLPC	50	39	41	39	46	40	42
GBC	74	62	43	51	54	51	45
MLPR	99	99	98	94	99	99	88
LASSO	99	99	98	93	99	99	85
DT	95	94	66	92	91	89	84
RF	97	95	87	96	96	97	96

Tabla 24. Evaluación del rendimiento del estimador del algoritmo anual set de datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
SVMR	28	34	88	23	13	34	43
SVMC	65	58	57	69	70	65	63
LR	63	30	16	46	66	38	38
LM	55	30	16	45	67	39	39
MLPC	61	35	43	53	72	20	40
GBC	96	65	43	96	100	82	84
MLPR	86	81	98	78	85	92	90
LASSO	99	98	98	99	99	99	90
DT	98	93	74	100	100	99	99
RF	97	94	90	99	99	98	99

En las tablas anteriores se puede evidenciar que los cuatro últimos algoritmos de cada tabla, cuentan con un rendimiento del estimador casi en todos los casos de entre el 80% y el 100% para casi todos los periodos anuales, los algoritmos que cuentan con varios

datos exacos del 100% o cercanos al 100%, no son una buena elección ya que el modelo podría sufrir de sobre entrenamiento y puede estar forzado lo cual puede ser perjudicial para el resto del análisis, por esa razón ninguno es descartado para el análisis posterior y los algoritmos SVMR, SVMC, LR, LM, MLPC y GBC tienen valores inferiores al 75% y en otros casos hasta valores menores en unos de los registros anuales o en todos los registros anuales, eso refleja que las clases que tiene el set de datos actual son muy pocas para que el algoritmo pueda clasificar la información con mejor rendimiento y de éstos el algoritmo SVMC es el que más claro muestra que si existirán otras clases es muy posible que el algoritmo fuera ideal para el análisis debido a que en sus registros anuales sus diferencias no son mayores a siete puntos porcentuales en el set de datos de confirmados y no mayor a tres puntos porcentuales en el set de datos completo, por eso quedan descartados como opciones. Entonces los candidatos a evaluar la información son los algoritmos MLPR, LASSO, RF y DT.

VI. ALGORITMO DE IDENTIFICACIÓN DE CARACTERÍSTICAS PARA EL SET DE DATOS.

Dado que la correlación de los datos con la clase género no presenta dependencia, se tendrá excluida esta variable en los siguientes análisis que determinan la importancia de cada clase en el set de datos. Al tener conocimiento de la información con que cuenta el set de datos a estudiar con machine learning, se hace necesario elegir un algoritmo que permita estudiar el fenómeno con un buen porcentaje de acierto en el aprendizaje de clasificador a usar, por este motivo se pondrán a estudio los algoritmos que se muestran en la siguiente tabla 25.

Tabla 25. Algoritmos de Identificación de Características.

NÚMERO	ALGORITMO	SÍMBOLO
1	Random Forest Classifier	RFC
2	Random Forest Regressor	RFR
3	Extra Trees Classifier	ETC
4	ExtraTrees Regressor	ETR
5	Ada Boost Classifier	ABC
6	Ada Boost Regressor	ABR

El análisis de importancia de cada una de las características, por cada uno de los algoritmos propuestos en la tabla 23 se muestra en las ilustraciones de la 63 a la 73.

Ilustración 62. Características Principales RFC Datos Completos.

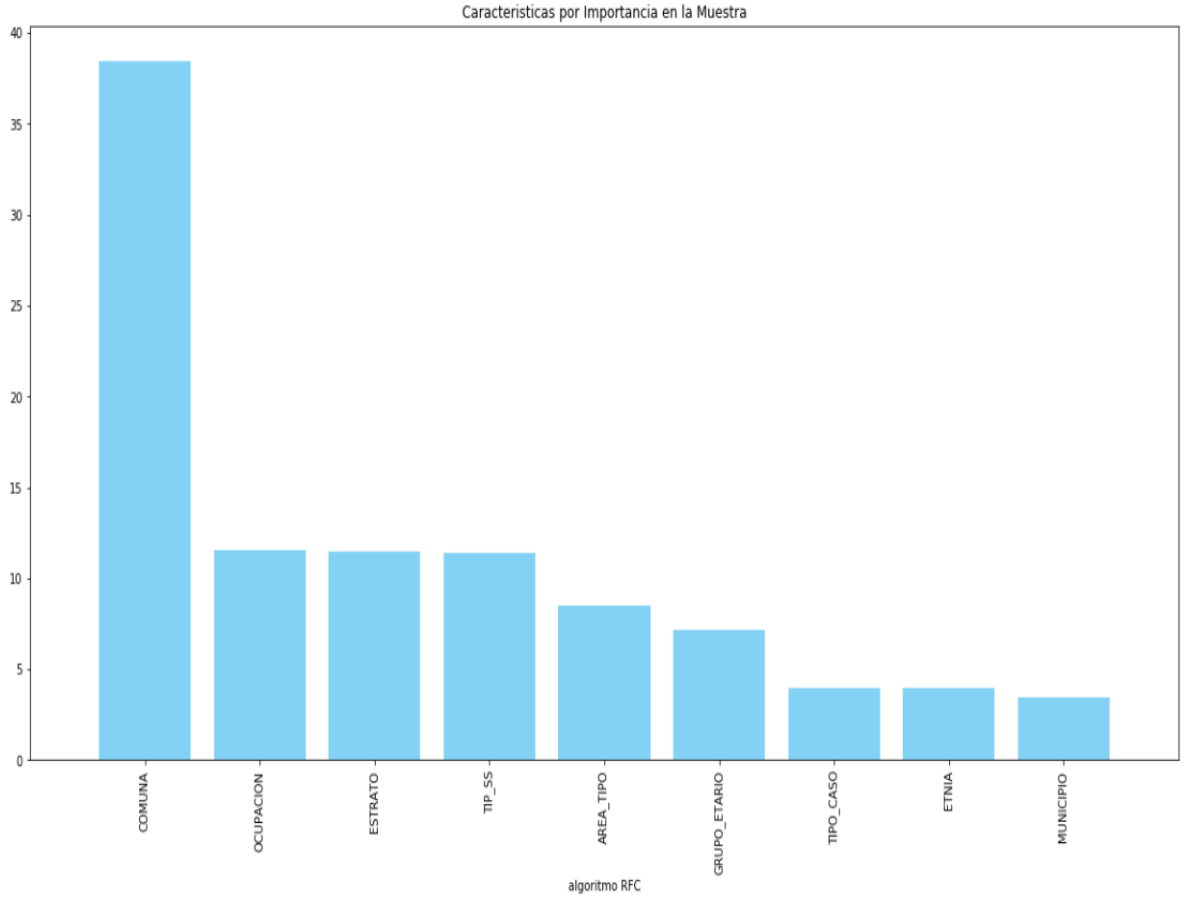


Ilustración 63. Características Principales RFC Datos Confirmados.

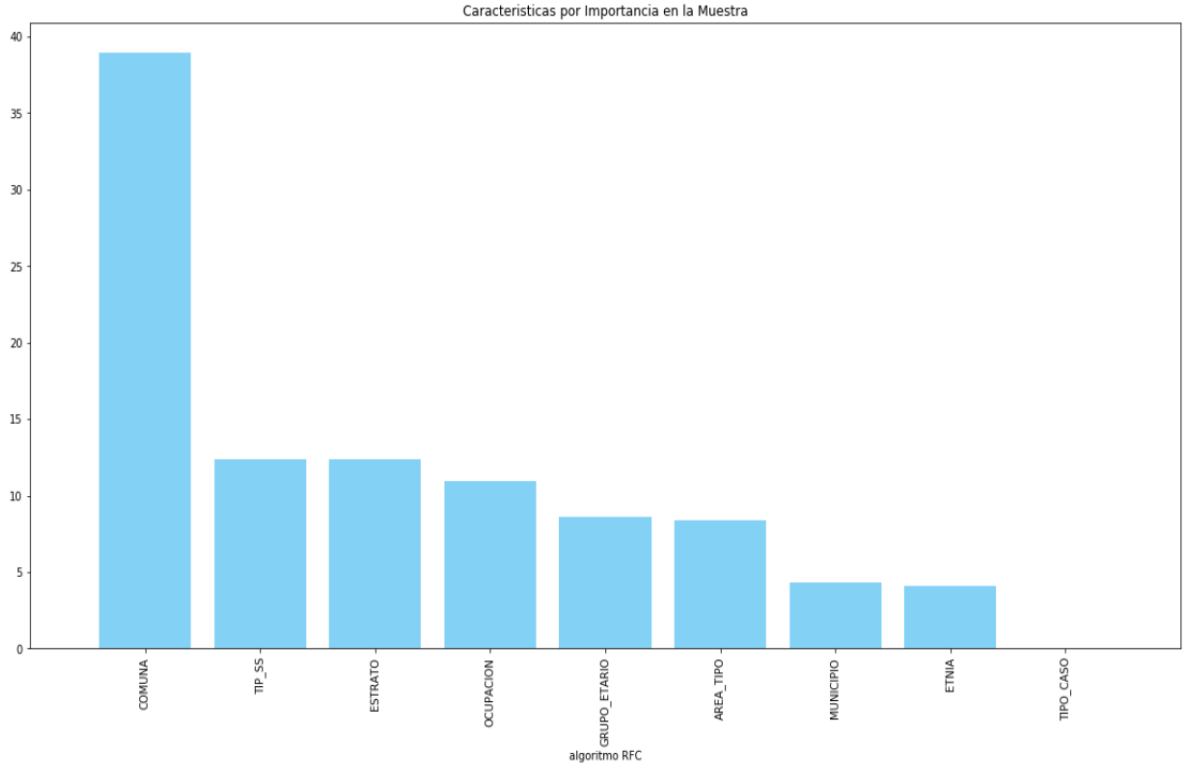


Ilustración 64. Características Principales RFR Datos Completos.

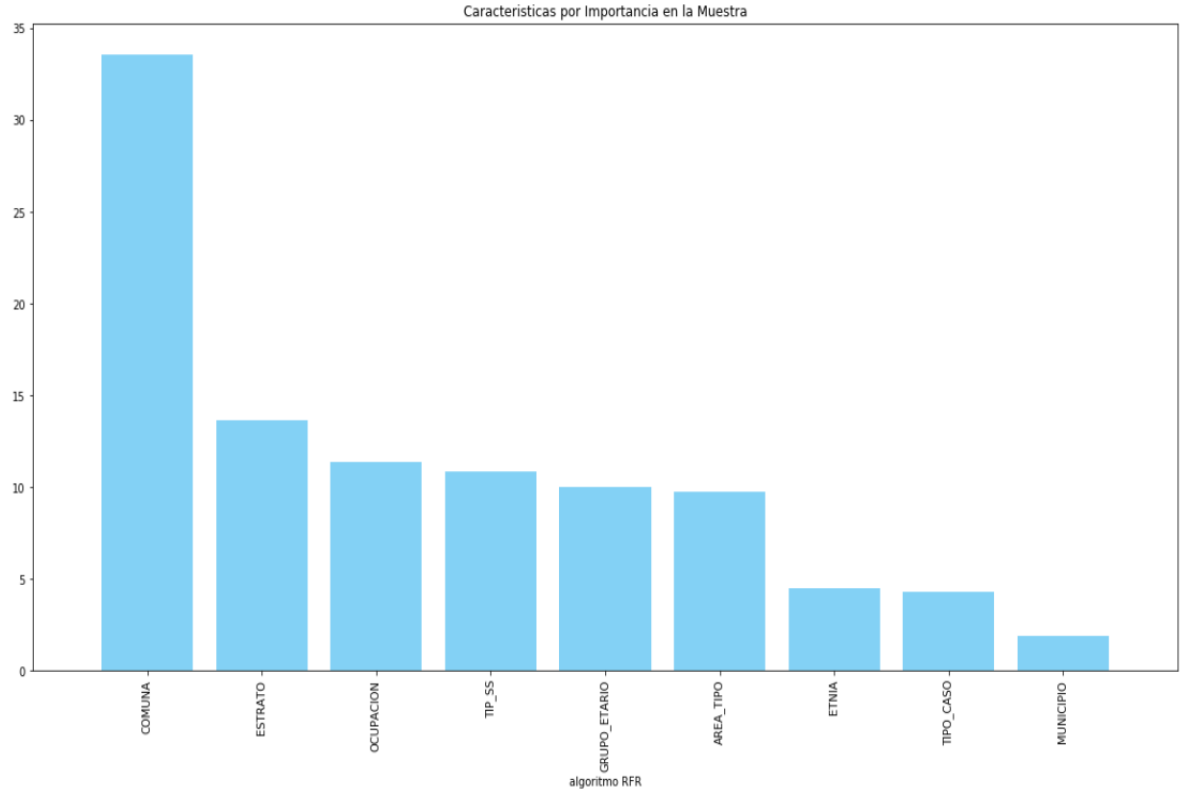


Ilustración 65. Características Principales RFR Datos Confirmados.

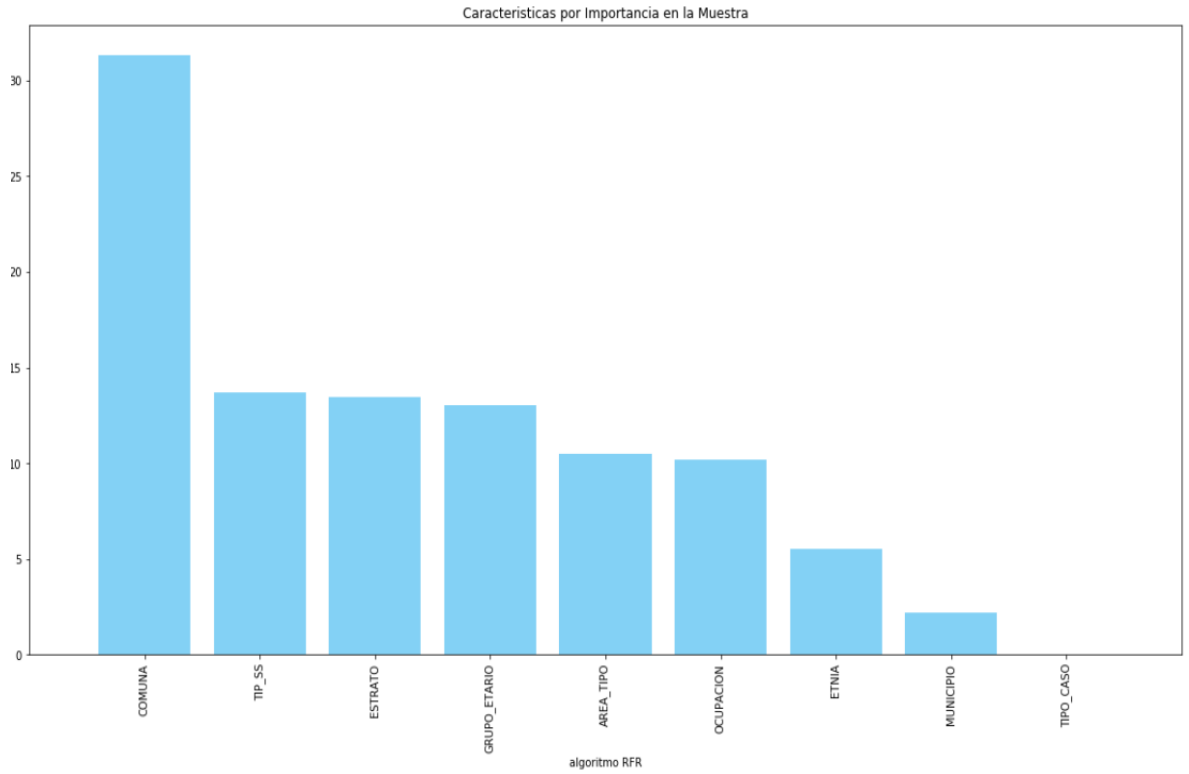


Ilustración 66. Características Principales ETC Datos Completos.

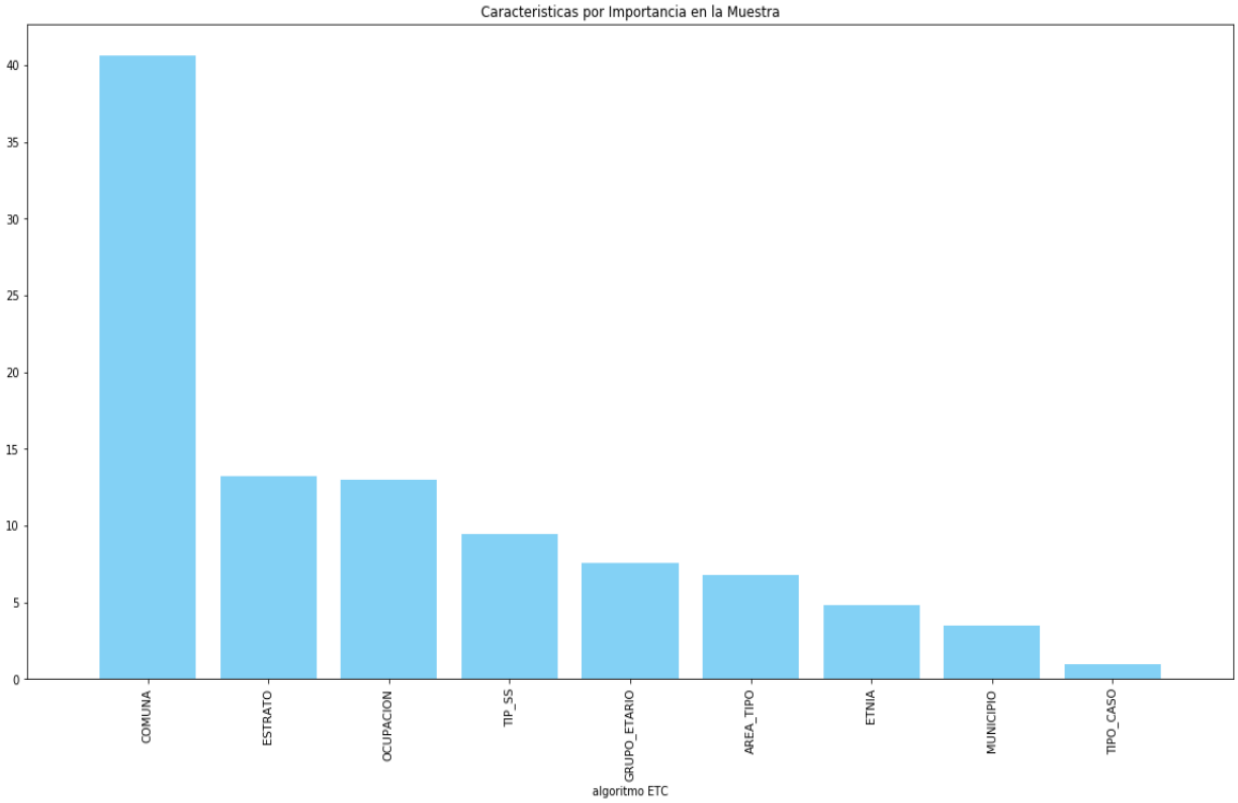


Ilustración 67.. Características Principales ETC Datos Confirmados.

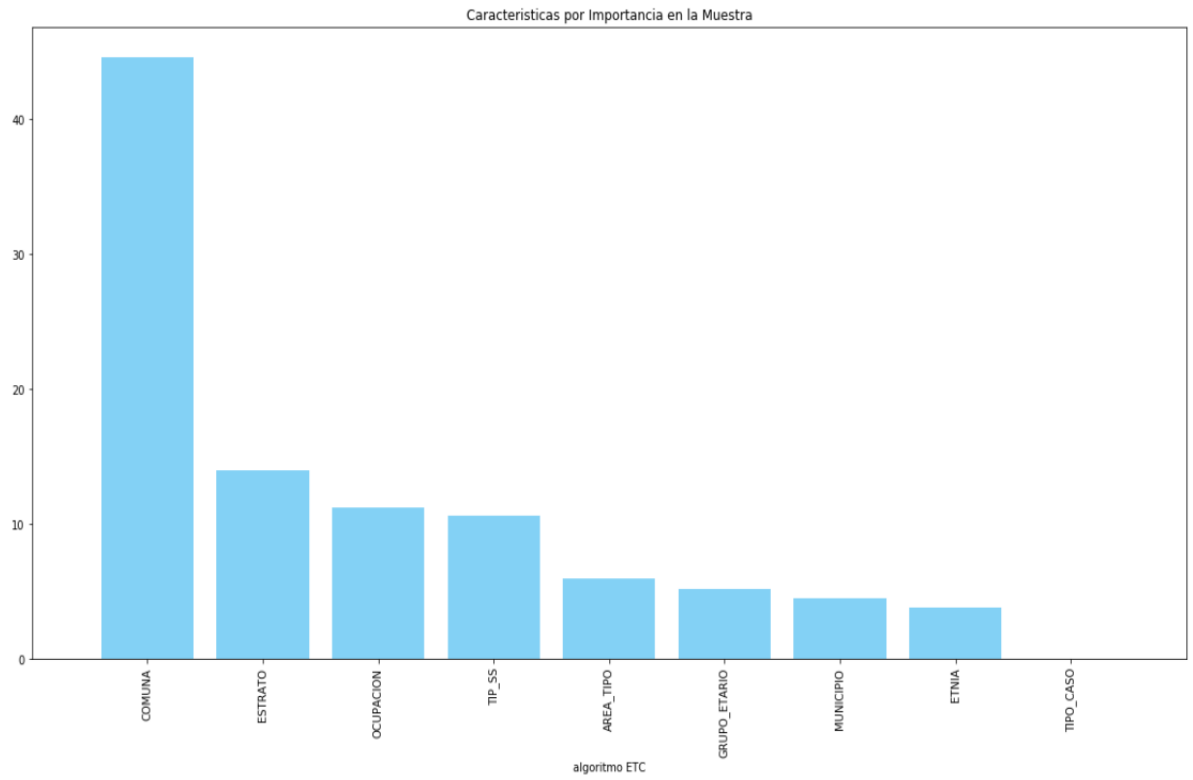


Ilustración 68. Características Principales ETR Datos Completos.

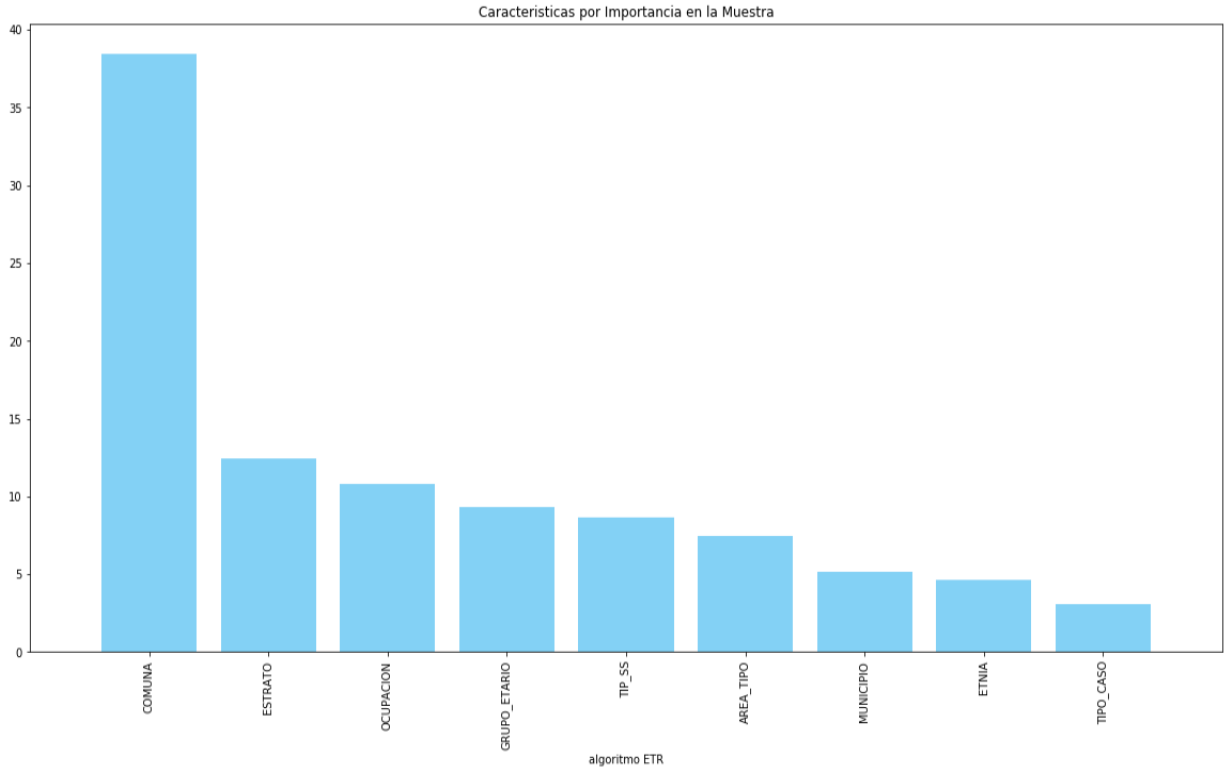


Ilustración 69. Características Principales ETR Datos Confirmados.

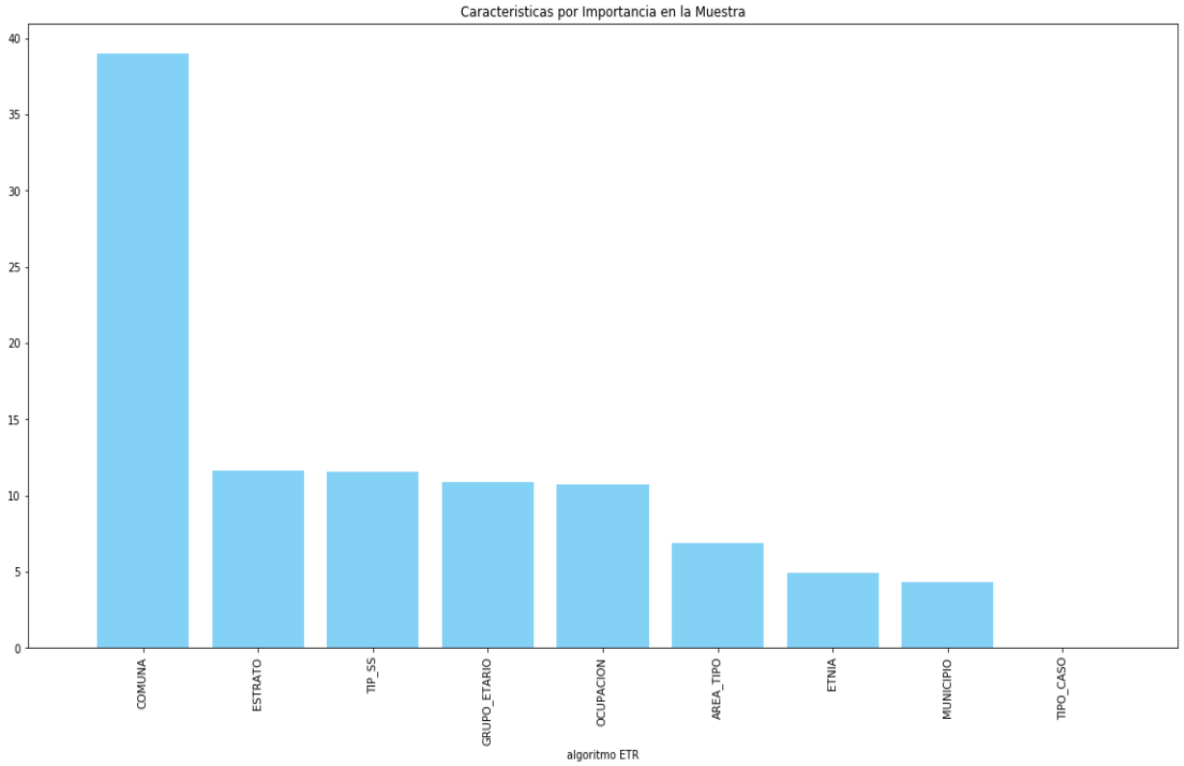


Ilustración 70. Características Principales ABC Datos Completos.

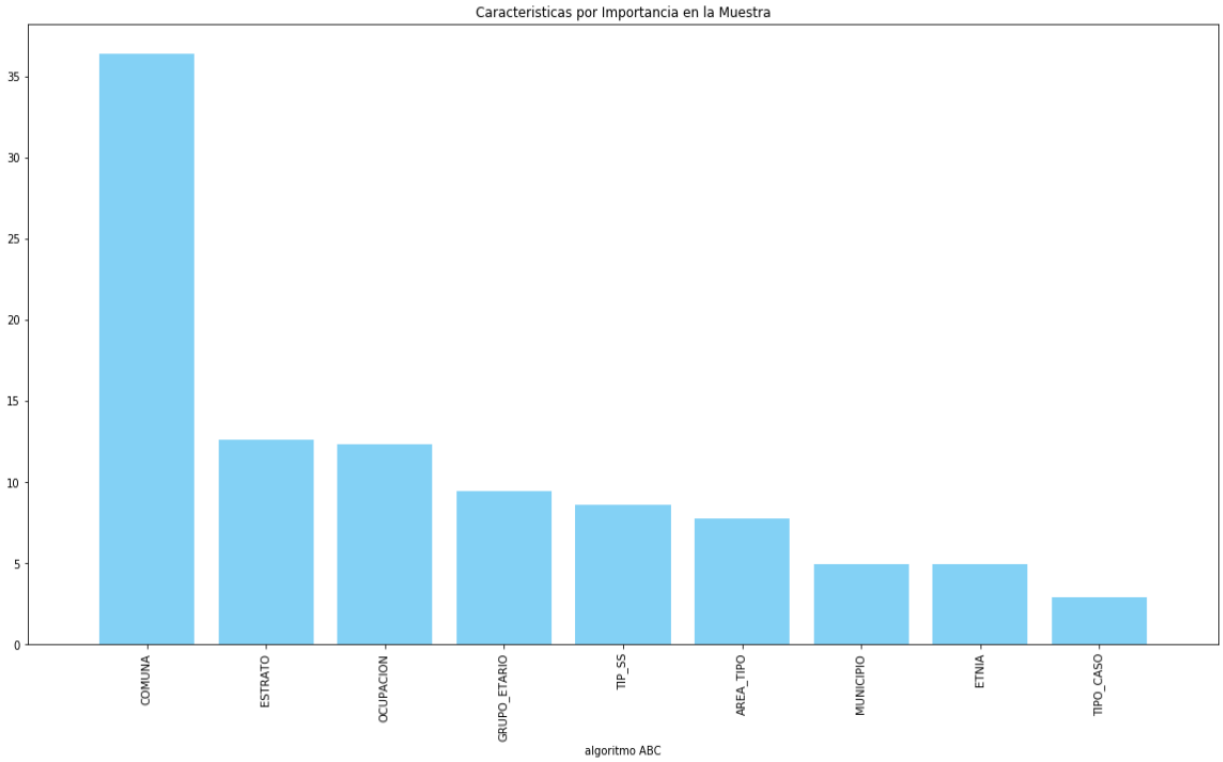


Ilustración 71. Características Principales ABC Datos Confirmados.

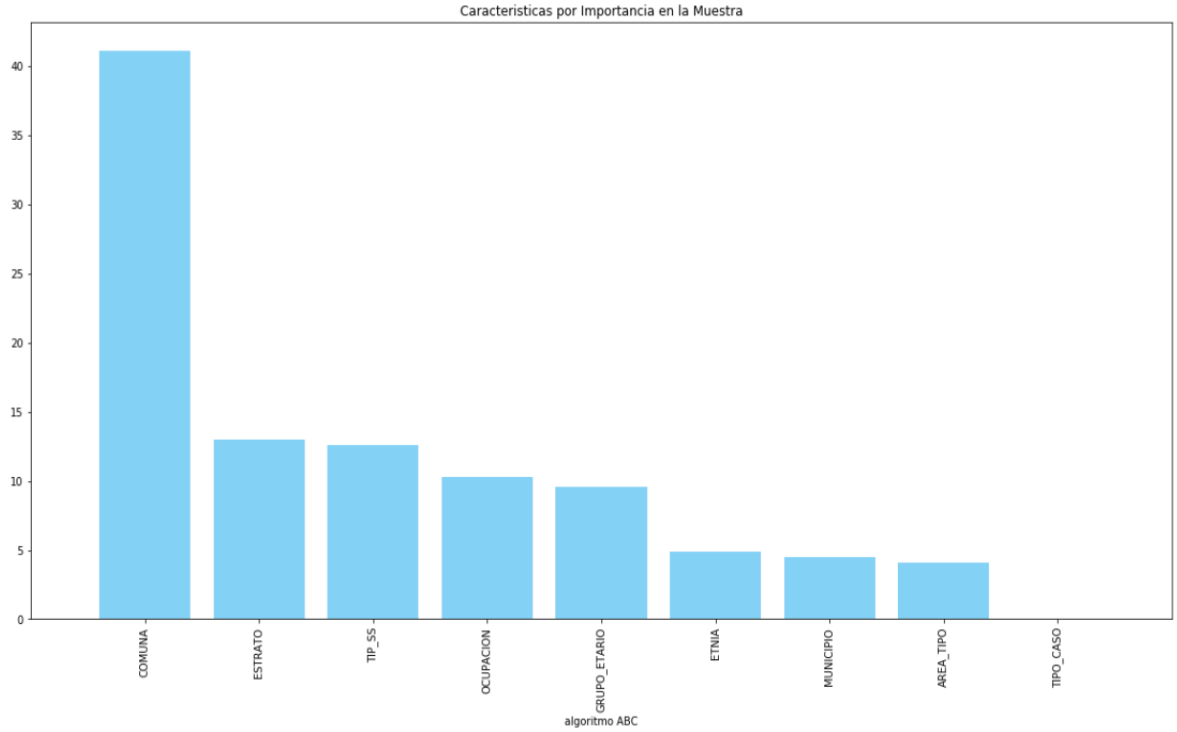


Ilustración 72. Características Principales ABR Datos Completos.

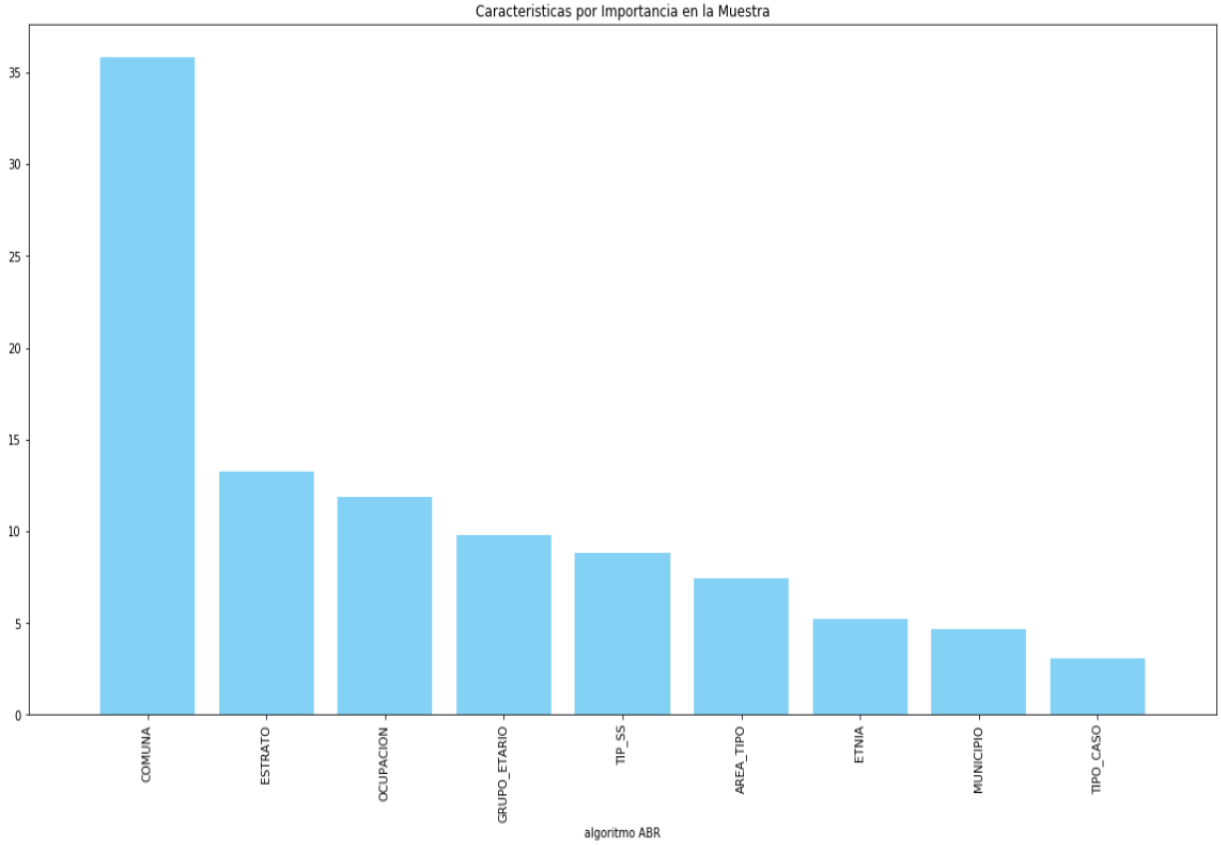


Ilustración 73. Características Principales ABR Datos Confirmados.

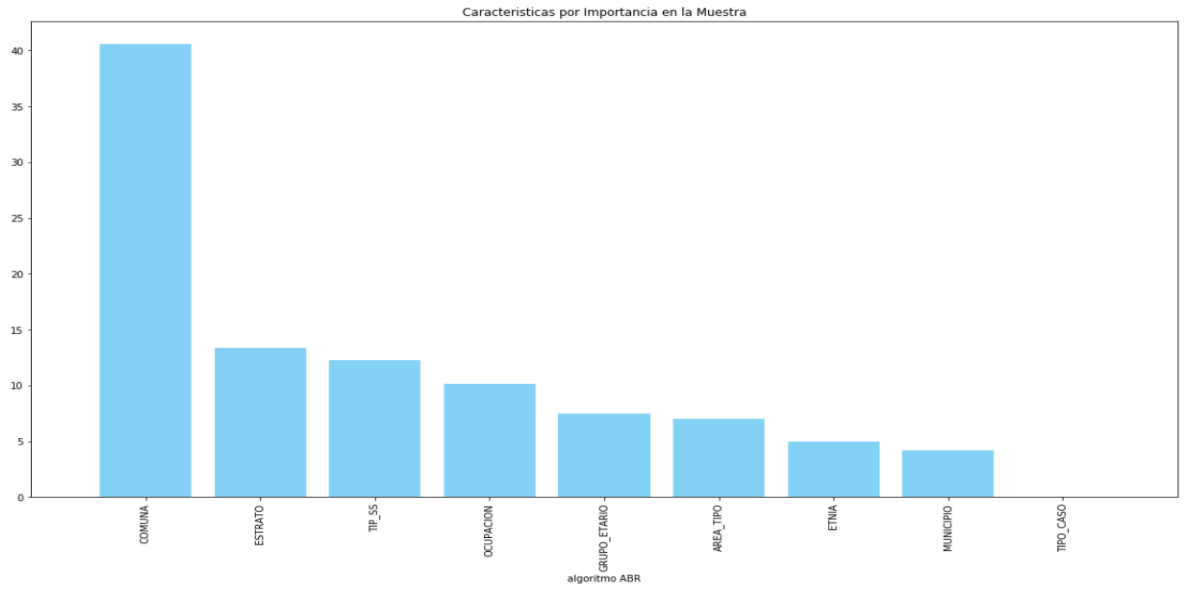


Tabla 26. Características Clases Principales Datos Completos.

CLASE	RFC	RFR	ETC	ETR	ABC	ABR	PROMEDIO
GRUPO ETARIO	7.2	10.1	7.6	9.3	9.4	9.8	8.9
ÁREA TIPO	8.5	9.8	6.8	7.4	7.8	7.5	8.0
OCUPACIÓN	11.6	11.4	13.0	10.8	12.4	11.9	11.9
TIPO SS	11.4	10.9	9.5	8.7	8.6	8.8	9.7
ESTRATO	11.5	13.6	13.3	12.4	12.6	13.3	12.8
ETNIA	4.0	4.5	4.8	4.6	4.9	5.2	4.7
MUNICIPIO	3.4	1.9	3.5	5.2	5.0	4.7	4.0
COMUNA	38.5	33.6	40.6	38.5	36.4	35.8	37.2
TIPO CASO	4.0	4.3	1.0	3.1	2.9	3.0	3.1

Tabla 27. Características Clases Principales Datos Confirmado.

CLASE	RFC	RFR	ETC	ETR	ABC	ABR	PROMEDIO
GRUPO ETARIO	8.6	13.0	5.2	10.9	9.6	7.5	9.1
ÁREA TIPO	8.4	10.5	6.0	6.9	4.1	7.0	7.2
OCUPACIÓN	10.9	10.2	11.2	10.7	10.3	10.1	10.6
TIPO SS	12.4	13.7	10.7	11.6	12.6	12.3	12.2
ESTRATO	12.4	13.5	14.0	11.7	13.0	13.4	13.0
ETNIA	4.1	5.5	3.8	4.9	4.9	5	4.7
MUNICIPIO	4.3	2.2	4.5	4.3	4.5	4.2	4.0
COMUNA	39	31.4	44.6	39.0	41.2	40.6	39.3
TIPO CASO	0	0	0	0	0	0	0.0

Las tablas anteriores 26 y 27, muestran las características importantes del modelo de datos en formato porcentual, tanto para el set de datos completo como para el set de datos confirmado, al realizar el análisis con los seis algoritmos propuestos y medir el promedio entre ellos se da una medida aceptable de cada una de las clases que componen el modelo, esto permite determinar la relación que tienen las clases con la efectividad del clasificador con respecto al modelo. Por lo tanto, dice en cada clase la relación de éstas con el éxito de la predicción del modelo cuando se tenga en cuenta para la evaluación, donde en casi todos los casos la clase comuna consigue una dependencia de la tercera parte de la importancia en el modelo, la parte económica que va relacionado con la ocupación de la persona representa algo por encima del diez por ciento y la parte social que está ligada con la seguridad social y el estrato de la persona explican en promedio más del veintitrés por ciento del modelo, pero esto será verificado con los algoritmos seleccionados y se podrá evidenciar que tanto afectan el clasificador en su nivel de precisión del modelo en forma negativa al estar ausente alguna de estas variables.

VII. ALGORITMOS FACTIBLES PARA EL SET DE DATOS

Anteriormente se evaluaron los algoritmos de machine learning y fueron elegidos como probables el MLPR, LASSO, RF Y DT, se realizarán las evaluaciones del set de datos con estos cuatro algoritmos y se evaluará el cambio en la precisión de los modelos. Veamos cómo se comportan los algoritmos al eliminar la clase mes con las ilustraciones de la 74 al 79.

Ilustración 74. Evaluación Algoritmo MLPR Sin la clase MES.

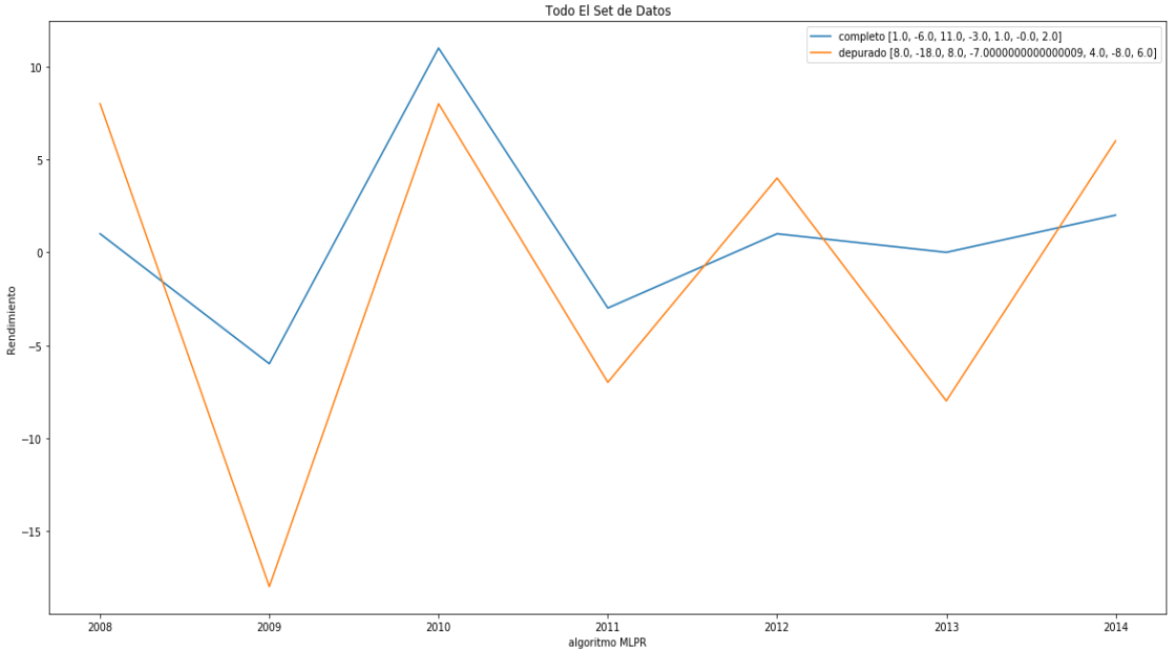


Ilustración 75. Evaluación Algoritmo LASSO Sin la clase MES.

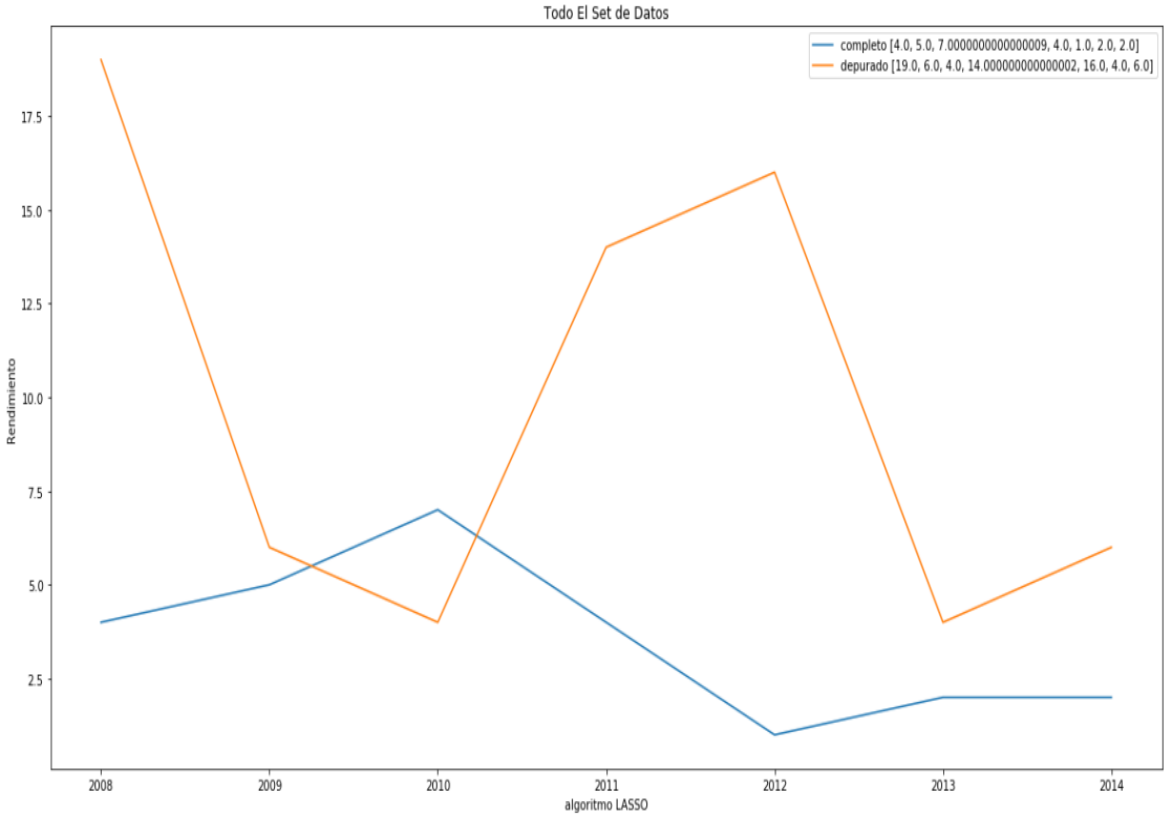


Ilustración 76. Evaluación Algoritmo RF Sin la clase MES.

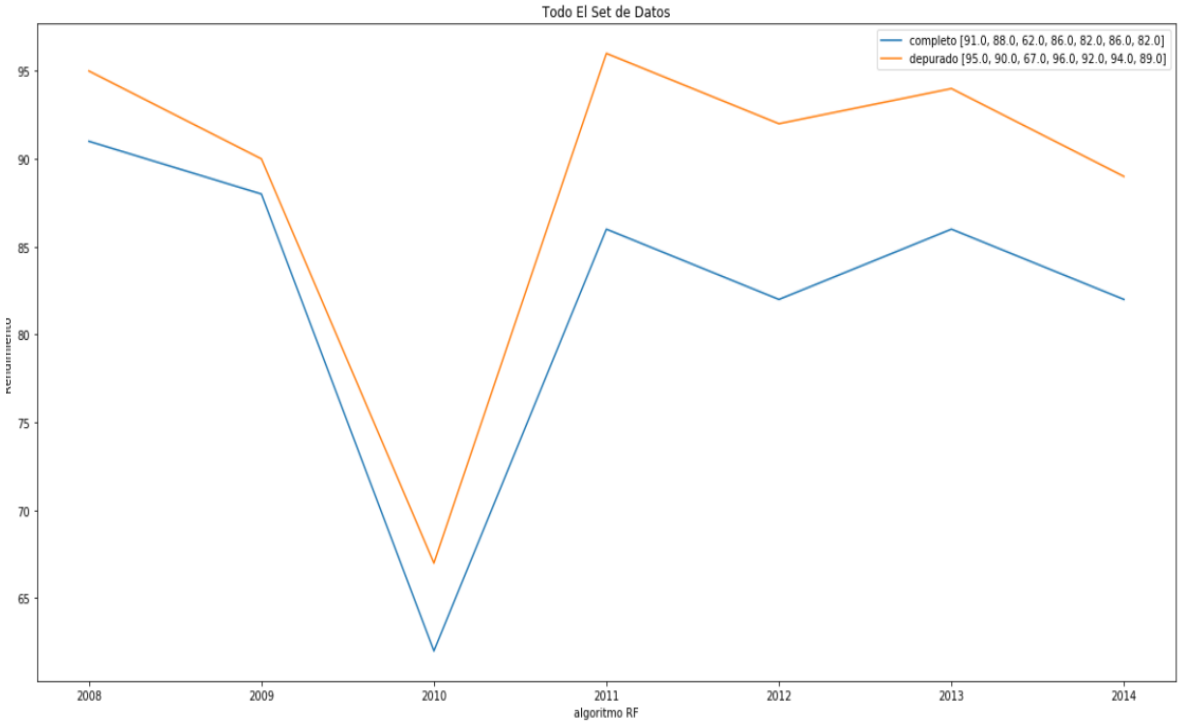


Ilustración 77. Evaluación Algoritmo DT Sin la clase MES.

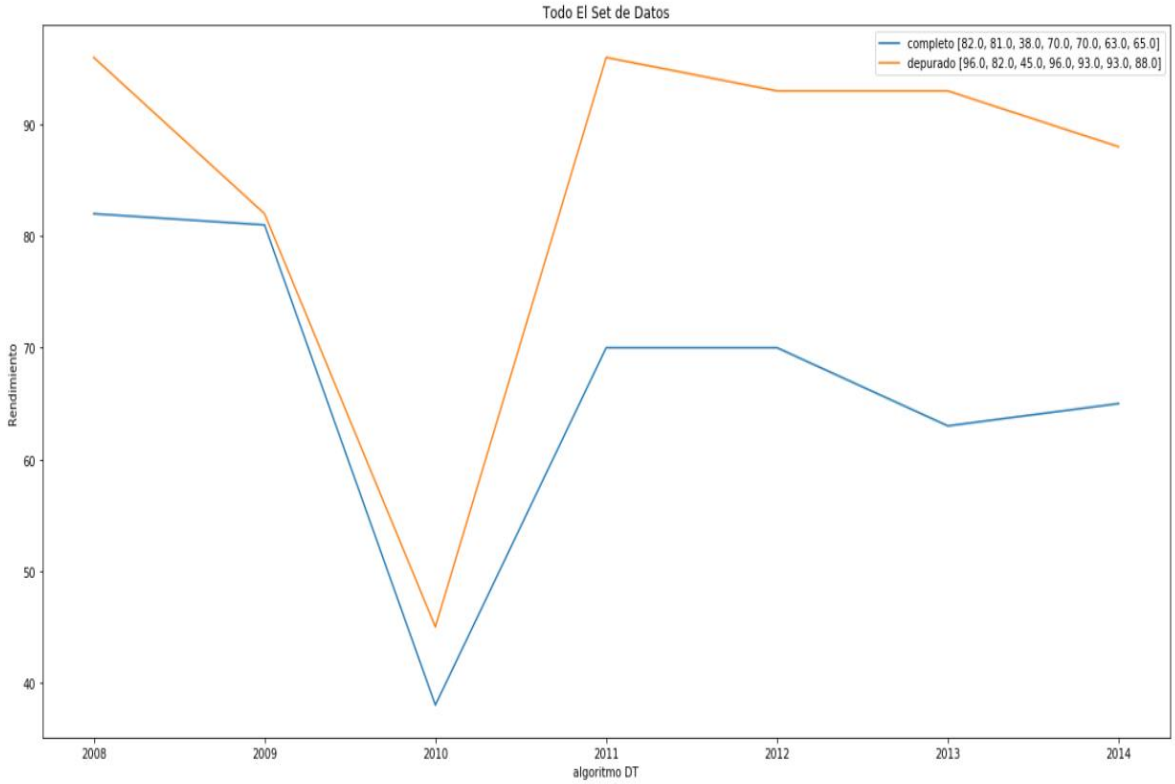


Ilustración 78. Resumen Algoritmos Sin la clase MES Set de Datos Completo.

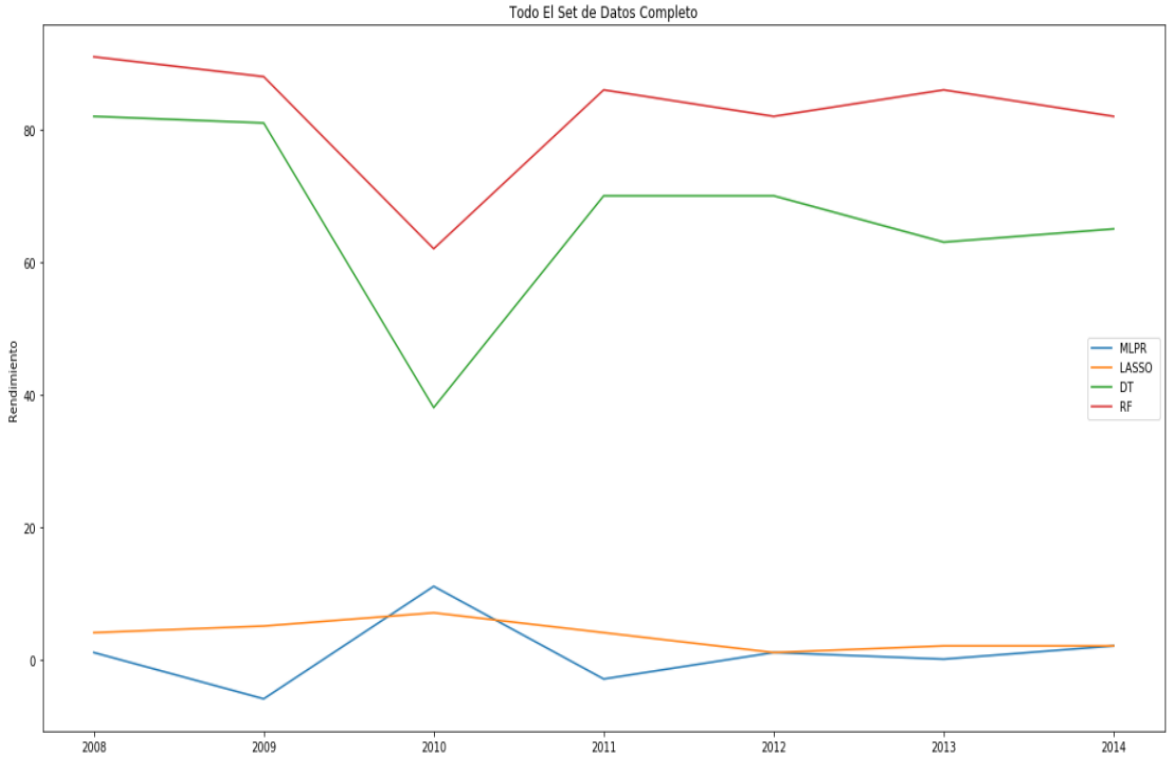


Ilustración 79. Resumen Algoritmos Sin la clase MES Set de Datos Confirmados.

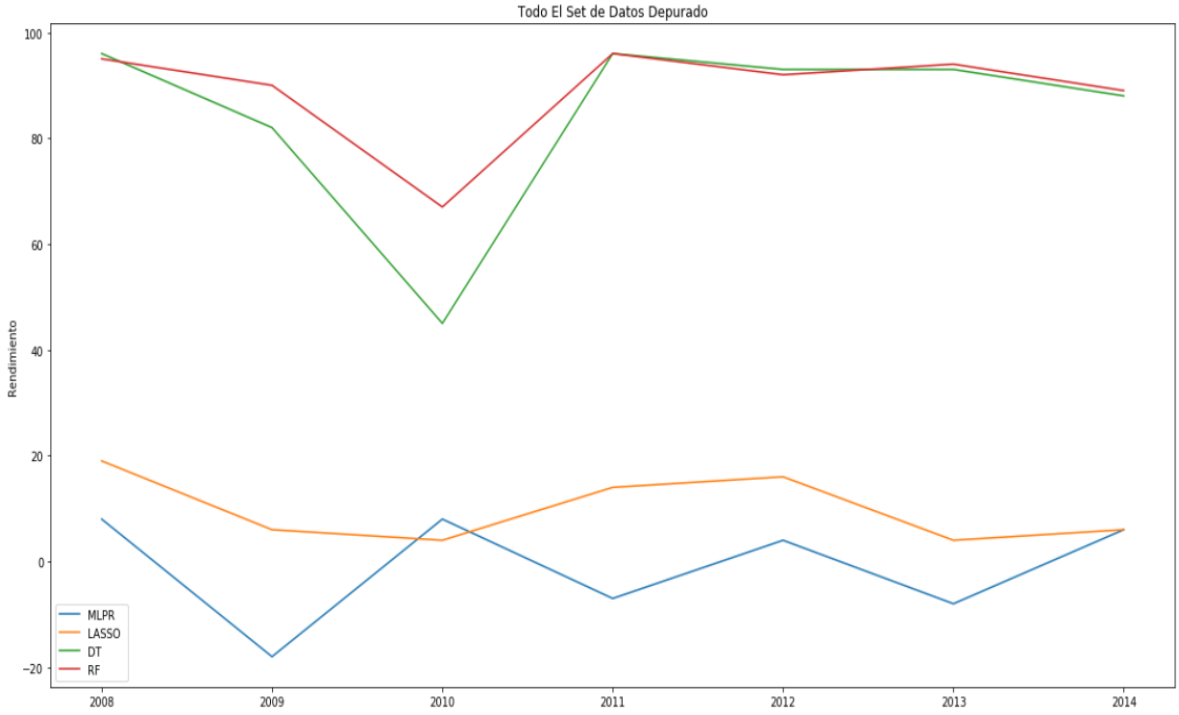


Tabla 28. Comparación anual rendimiento algoritmos seleccionados datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
MLPR	1.0	-6.0	11.0	-3.0	1.0	0.0	2.0
LASSO	4.0	5.0	7.0	1.0	2.0	2.0	2.0
DT	82.0	81.0	38.0	70.0	70.0	63.0	65.0
RF	91.0	88.0	62.0	86.0	82.0	86.0	82.0

Tabla 29. Comparación anual rendimiento algoritmos seleccionados datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
MLPR	8.0	-18.0	8.0	-7.0	4.0	-8.0	6.0
LASSO	19.0	6.0	4.0	14.0	16.0	4.0	6.0
DT	96.0	82.0	45.0	96.0	93.0	93.0	88.0
RF	95.0	90.0	67.0	96.0	92.0	94.0	89.0

Con los resultados obtenidos en las tablas 28 y 29, y la clara diferencia en los gráficos de acierto en predicción, al eliminar la variable temporal MES, los algoritmos LASSO y MLPR quedan completamente descartados para el análisis dado que su tasa de aprendizaje es muy baja y no sería una buena decisión mantener este par de algoritmos con las clases que se tienen actualmente, también se observa que para la muestra anual referente al 2010 el nivel de aprendizaje bajó considerablemente, pero en los demás datos anuales el promedio de éxito es bastante bueno superando el 80% en casi todos los casos. Bajo esta premisa los algoritmos que seguirán en el análisis son RF y DT.

VIII. ALGORITMOS ELEGIDOS PARA ANÁLISIS EXPLICATIVO DEL SET DE DATOS

En los análisis que se realizaron anteriormente, donde se evaluaron cada una de las diez máquinas de aprendizaje propuestas con parámetros por defecto, poco a poco se fueron descartando aquellas máquinas en las cuales el nivel de predicción que conseguían se consideró bajo para el conjunto de datos analizados, llegando a una última instancia donde se tienen como factibles las máquinas referentes al Decision tree y Random Forest. Si se observan las máquinas que mejor resultados aportaron son las que están basadas en clasificadores tipo árbol, estos algoritmos particularmente utilizan regresión lo cual se basa en la relación de las variables de tipo categóricos e igualmente numéricas, y como anteriormente se demostró en el análisis relacional del set de datos, existe una fuerte relación entre muchas de las variables del set de datos que son de interés para este trabajo. Es un gran punto a tener en cuenta, se observa que la clase mes es numérica y el resto de las clases con que cuenta el set de datos son de tipo categórico y este par de clasificadores son muy utilizados como intérpretes de los datos evaluados con machine learning en tareas que tengan que ver con regresión y clasificación (Wiener, 2002), estos algoritmos aportan un resultado generalmente bueno para datos de alta dimensionalidad y son poco afectados por la profundidad con que los árboles cuenten, mientras que otro tipo de clasificadores son poco explicativos o no generan buen rendimiento cuando existen mucha diversidad de miembros dentro de una clase (Breiman, 2001). Esto puede explicar la disminución del rendimiento de los demás algoritmos con respecto al tipo árbol, Así que las máquinas de aprendizaje propuestas para continuar con el análisis para este set de datos son los Decision tree y Random Forest.

IX. EVALUACIÓN DE VARIABLES EXPLICATIVAS DEL MODELO

Ya se conoce con los puntos anteriores, que el set de datos se comporta de manera positiva frente a los clasificadores que usan regresión, entonces para encontrar un modelo que permita explicar de mejor manera la variación de alguna característica cuantificable es necesario tener una correcta evaluación y confirmación de un modelo conocido como modelo teórico, para este fin es necesario evaluar las variables que se tienen y determinar cuáles de ellas son tan determinantes que influyen en lo explicativo y predictivo que puede ser un modelo, Existen diferentes procedimientos que se usan para este tipo de tareas y son el método hacia atrás, eliminación regresiva o método backward, el método hacia delante, introducción progresiva o método forward y el método stepwise o regresión paso a paso (Ewout W. Steyerberg, 1999).

MÉTODO BACKWARD

Este método parte del principio de eliminar una a una las variables de menor a mayor importancia y cada vez que se elimina una variable se considera una etapa, hasta una regla de parada que regularmente es que la variable eliminada no modifique mucho la variabilidad de la predicción del modelo.

MÉTODO FORWARD

Este método parte del principio de adicionar una a una las variables de mayor a menor importancia y cada vez que se ingresa una variable se considera una etapa, hasta una

regla de parada que regularmente es que la variable adicionada no modifique mucho la variabilidad de la predicción del modelo.

MÉTODO STEPWISE

Este método parte del principio combinado de los métodos anteriores, comenzando por el método forward, pero cada vez que se evalúa una variable esta se mantiene en la muestra y se adiciona otra, con lo cual se construye paso a paso la variabilidad el modelo.

Apoyados en la sección algoritmos de identificación de características del set de datos, donde se calcula la importancia de cada clase, se tomarán estos análisis como punto de partida, de manera experimental se eliminará una sola clase y se revisará como afecta al modelo la falta de esta clase. Posteriormente se utilizará el método clásico (método backward) para evaluar las variables explicativas ya combinación los demás métodos, finalmente por último se hará un proceso inverso al backward para comparar la caída porcentual del modelo, eliminando variables de la más importante a la menos importante, en la siguiente tabla se muestran las siete clases que de mayor aportan al modelo.

Tabla 30. Resumen Importancia porcentual Clases del modelo.

IMPORTANCIA	CLASE	PROMEDIO
1	COMUNA	38.3
2	ESTRATO	12.9
3	OCUPACIÓN	11.2
4	TIPO SS	11
5	GRUPO ETARIO	9
6	ÁREA TIPO	7.6
7	ETNIA	4.7

ELIMINACIÓN EXPERIMENTAL DE UNA SOLA VARIABLE CON REEMPLAZO

Este informa qué tan bueno es el modelo cuando hay ausencia de una sola de las variables del set de datos, para ambos sets de datos, entonces en las ilustraciones de la 80 a la 93 se puede observar cómo impacta directamente la ausencia de cada una de las variables de manera independiente en el porcentaje que arroja el clasificador.

Ilustración 80. Eliminación Experimental DT COMUNA.

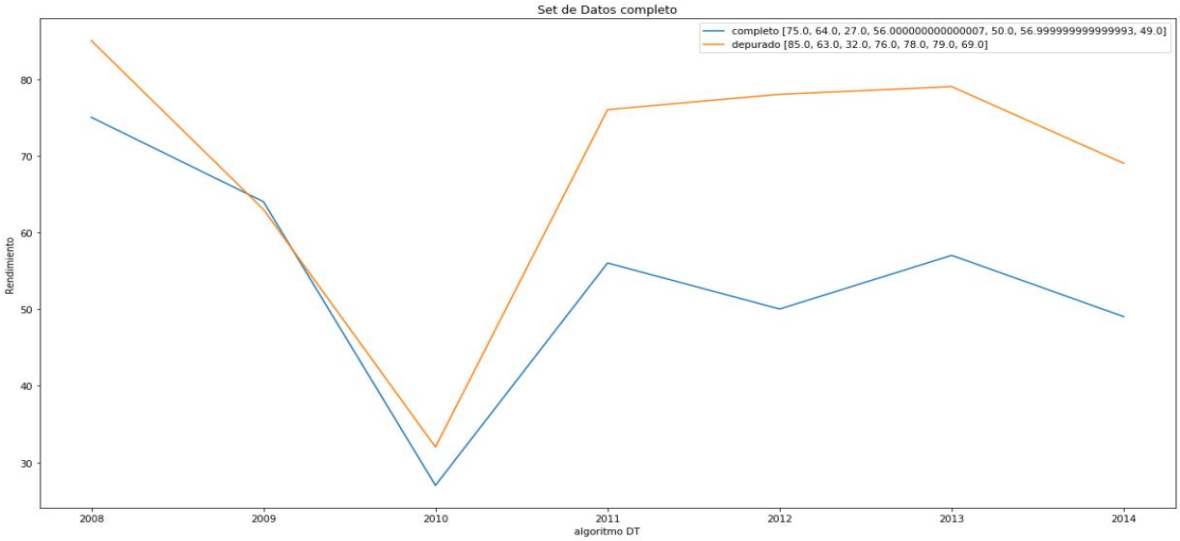


Ilustración 81. Eliminación Experimental RF COMUNA.

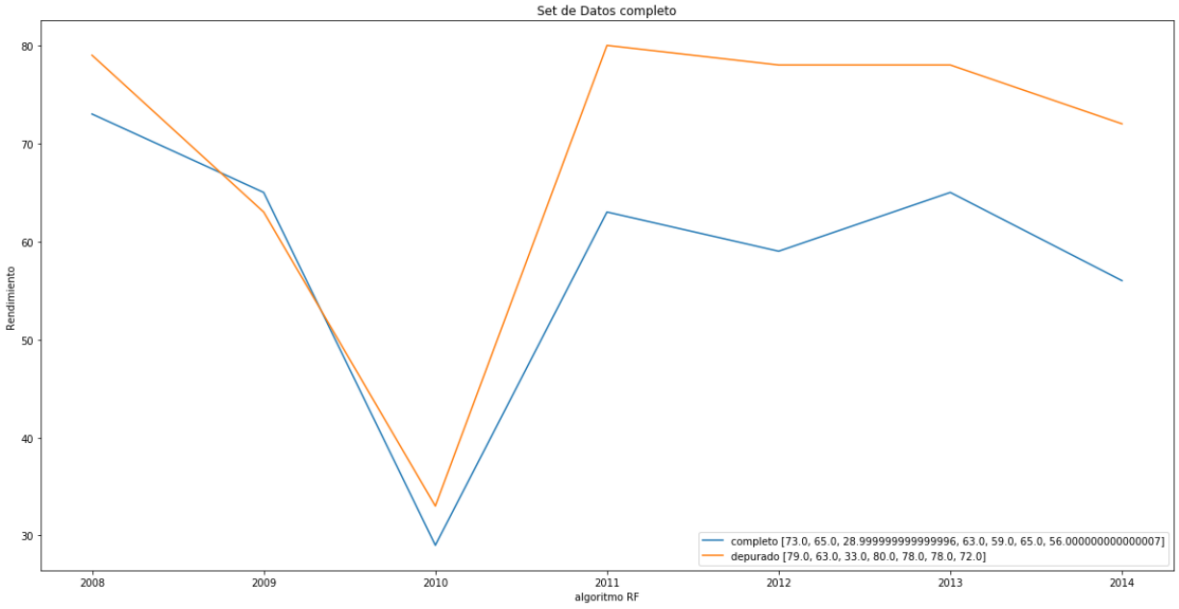


Tabla 31. Comparación anual rendimiento algoritmos seleccionados sin clase COMUNA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	75	64	27	56	50	57	49
RF	73	65	29	63	59	65	56

Tabla 32. Comparación anual rendimiento algoritmos seleccionados sin clase COMUNA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	85	63	32	76	78	79	69
RF	79	63	33	80	78	78	72

Ilustración 82. Eliminación Experimental RF ESTRATO.

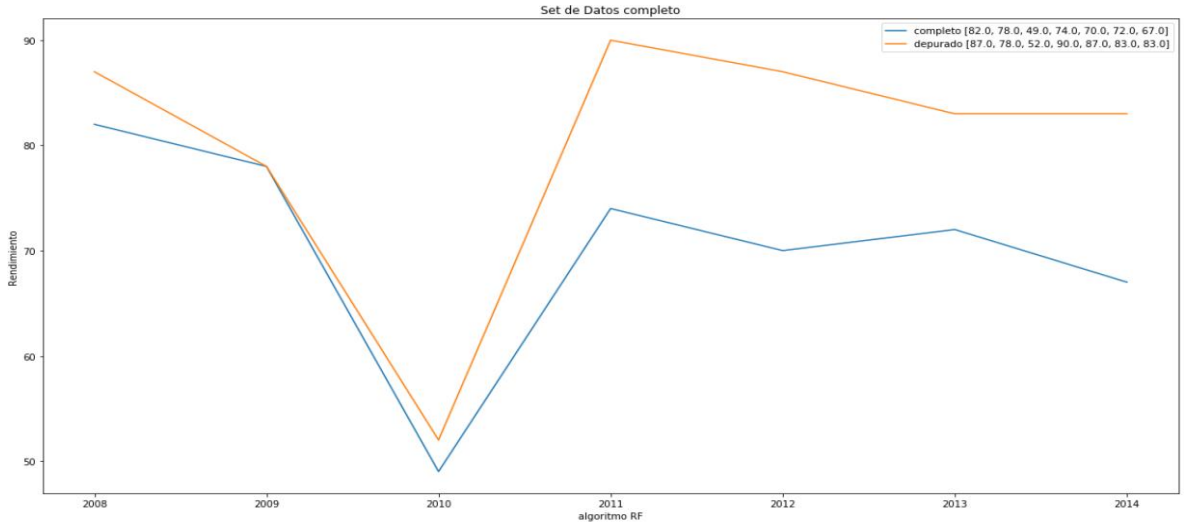


Ilustración 83. Eliminación Experimental DT ESTRATO.

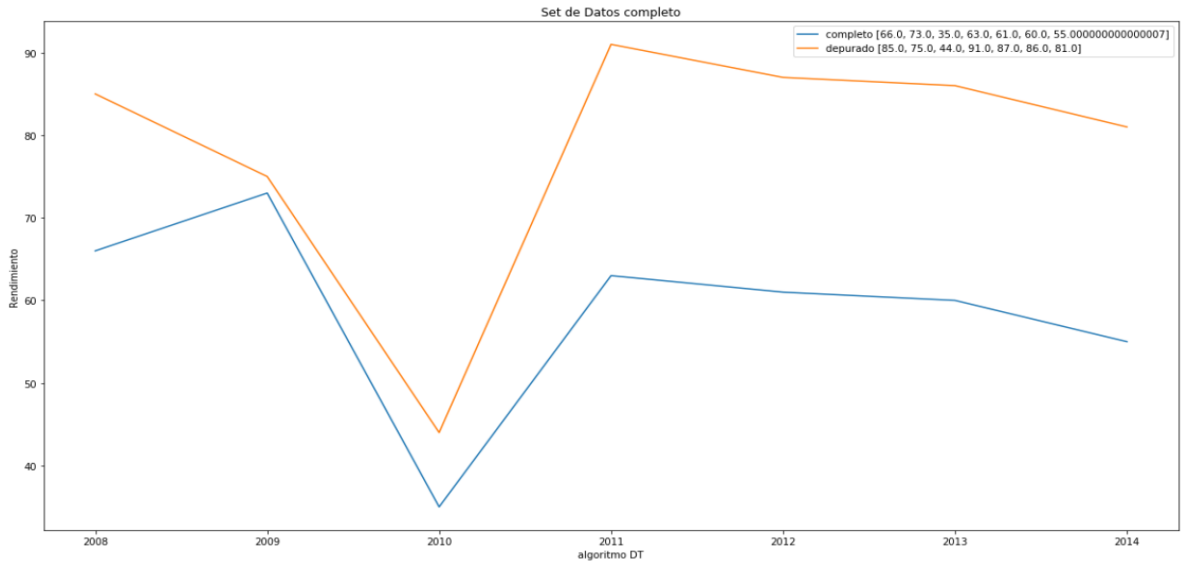


Tabla 33. Comparación anual rendimiento algoritmos seleccionados sin clase ESTRATO datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	66	73	35	63	61	60	55
RF	82	78	49	74	70	72	67

Tabla 34. Comparación anual rendimiento algoritmos seleccionados sin clase ESTRATO datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	85	75	44	91	87	86	81
RF	87	78	52	90	87	83	83

Ilustración 84. Eliminación Experimental DT OCUPACIÓN.

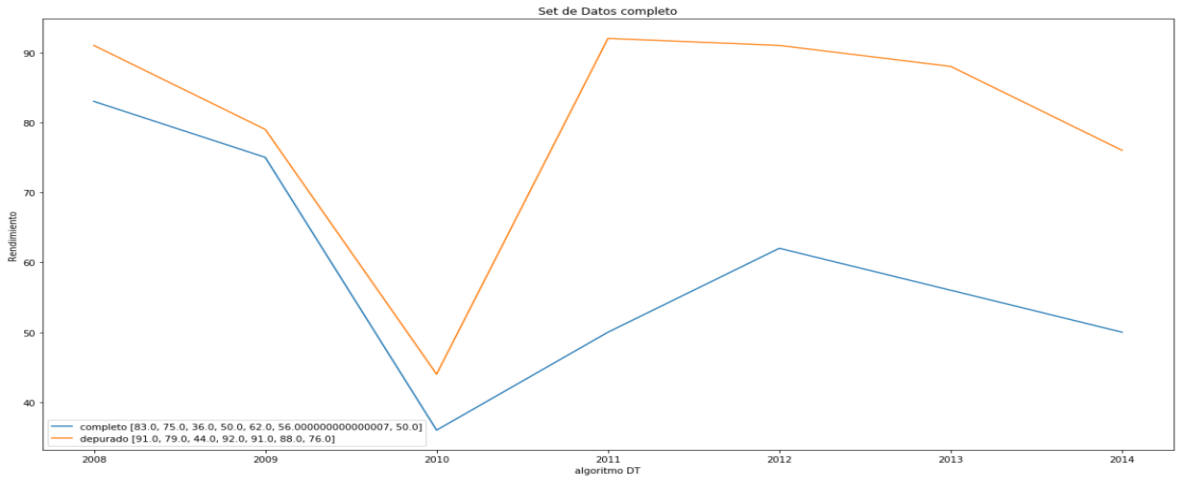


Ilustración 85. Eliminación Experimental RF OCUPACIÓN.

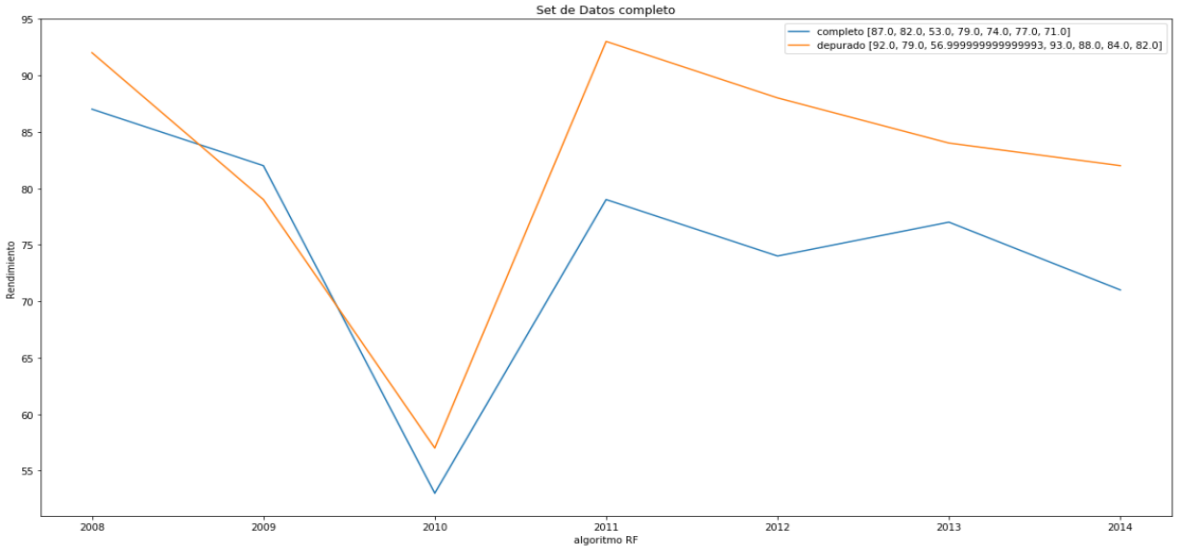


Tabla 35. Comparación anual rendimiento algoritmos seleccionados sin clase OCUPACIÓN datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	83	75	36	50	62	56	50
RF	87	82	53	79	74	77	71

Tabla 36. Comparación anual rendimiento algoritmos seleccionados sin clase OCUPACIÓN datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	91	79	44	92	91	88	76
RF	92	79	57	93	88	84	82

Ilustración 86. Eliminación Experimental RF ETNIA.

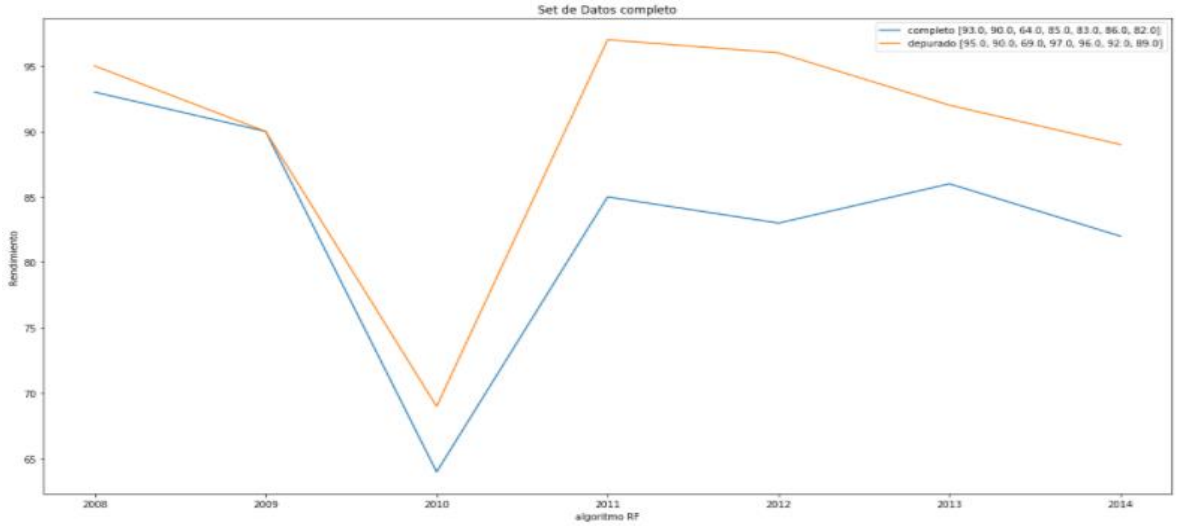


Ilustración 87. Eliminación Experimental DT ETNIA.

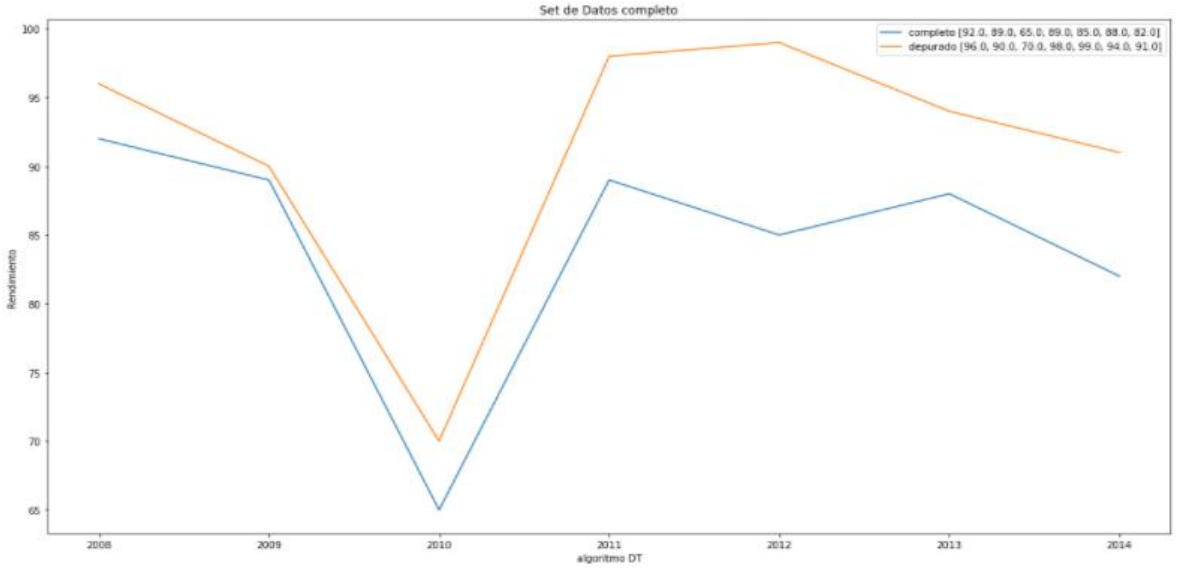


Tabla 37. Comparación anual Eliminación Experimental ETNIA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	92	89	65	89	85	88	82
RF	93	90	64	85	83	86	82

Tabla 38. Comparación anual Eliminación Experimental ETNIA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	96	90	70	98	99	94	91
RF	95	90	69	97	96	92	89

Ilustración 88. Eliminación Experimental RF GRUPO ETARIO.

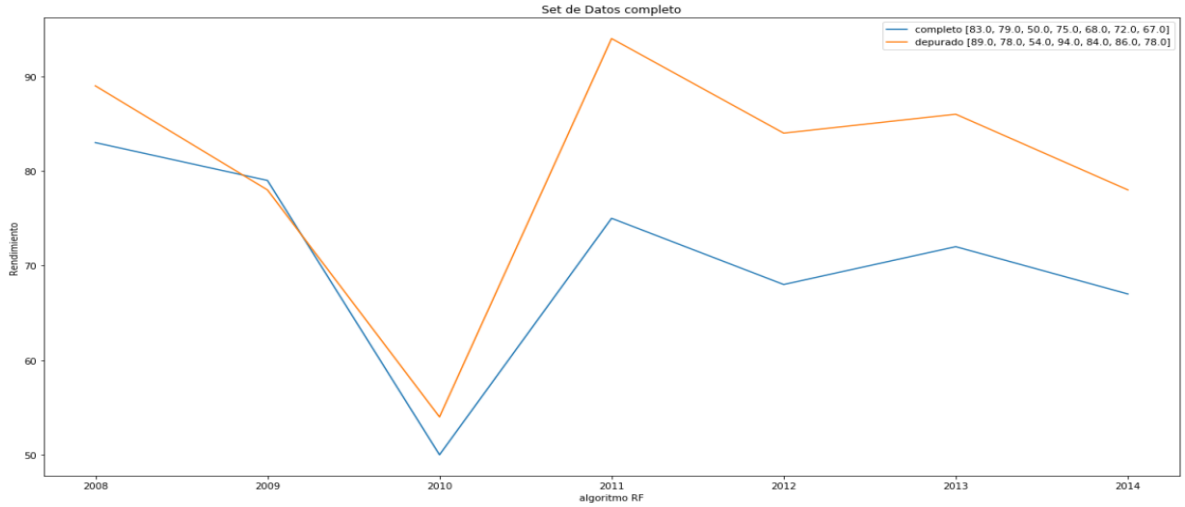


Ilustración 89. Eliminación Experimental DT GRUPO ETARIO.

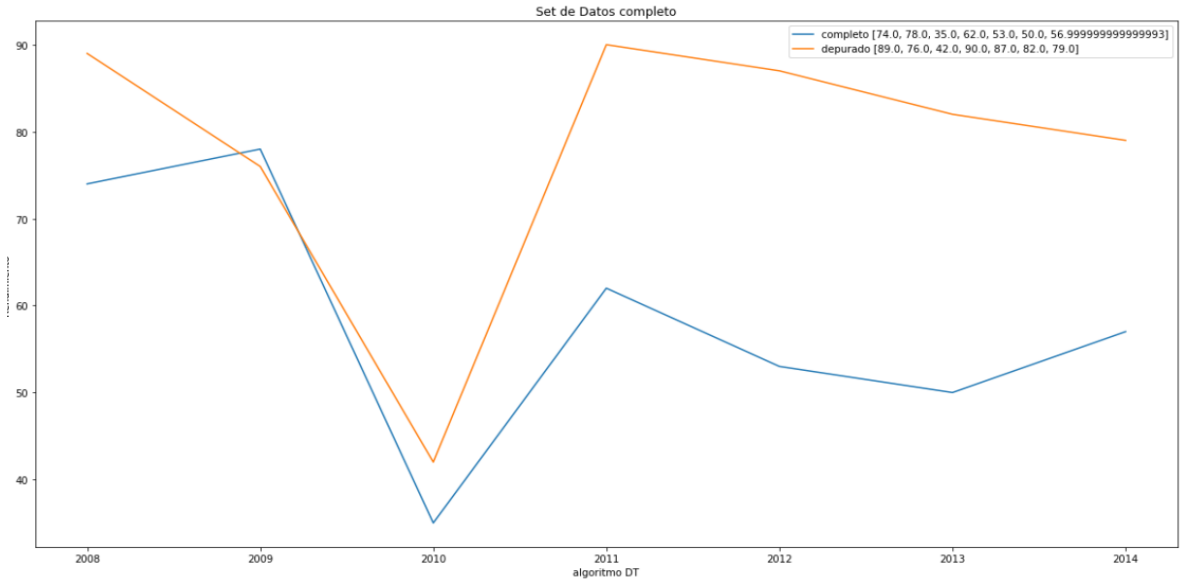


Tabla 39. Comparación anual rendimiento algoritmos seleccionados sin clase GRUPO ETARIO datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	74	78	35	62	53	50	57
RF	83	79	50	75	68	72	67

Tabla 40. Comparación anual rendimiento algoritmos seleccionados sin clase GRUPO ETARIO datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	89	76	42	90	87	82	79
RF	89	78	54	94	84	86	78

Ilustración 90. Eliminación Experimental DT ÁREA.

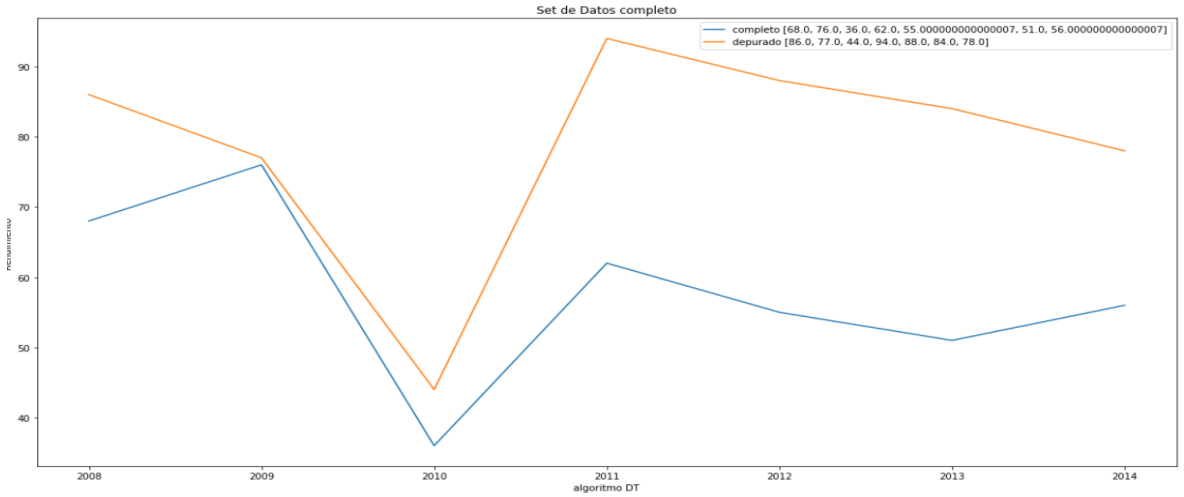


Ilustración 91. Eliminación Experimental RF ÁREA.

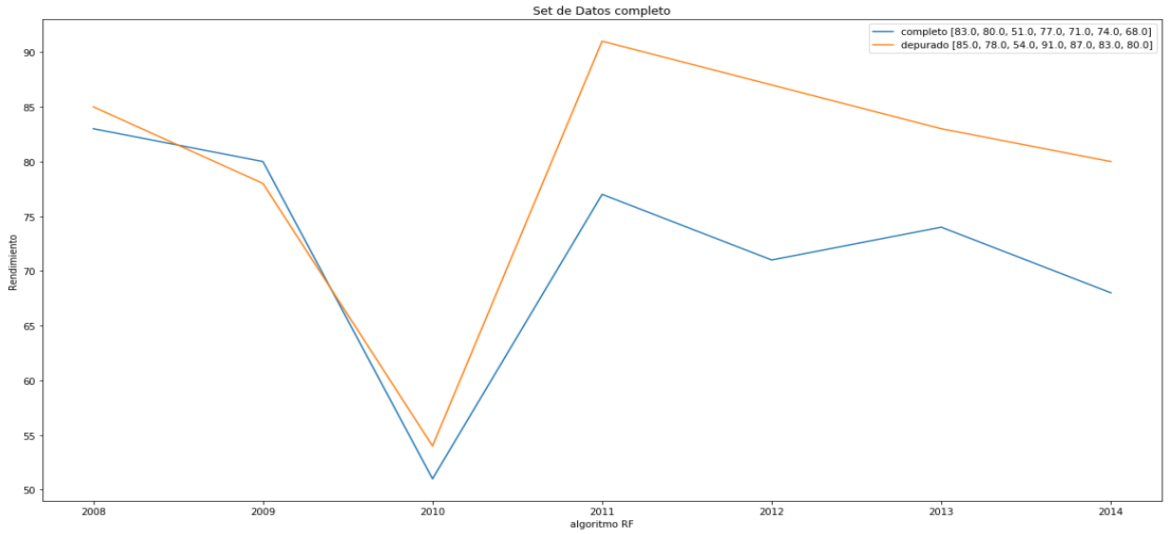


Tabla 41. Comparación anual rendimiento algoritmos seleccionados sin clase ÁREA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	68	76	36	62	55	51	56
RF	83	80	51	77	71	74	68

Tabla 42. Comparación anual rendimiento algoritmos seleccionados sin clase ÁREA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	86	77	44	94	88	84	78
RF	85	78	54	91	87	83	80

Ilustración 92. Eliminación Experimental DT SEGURIDAD SOCIAL.

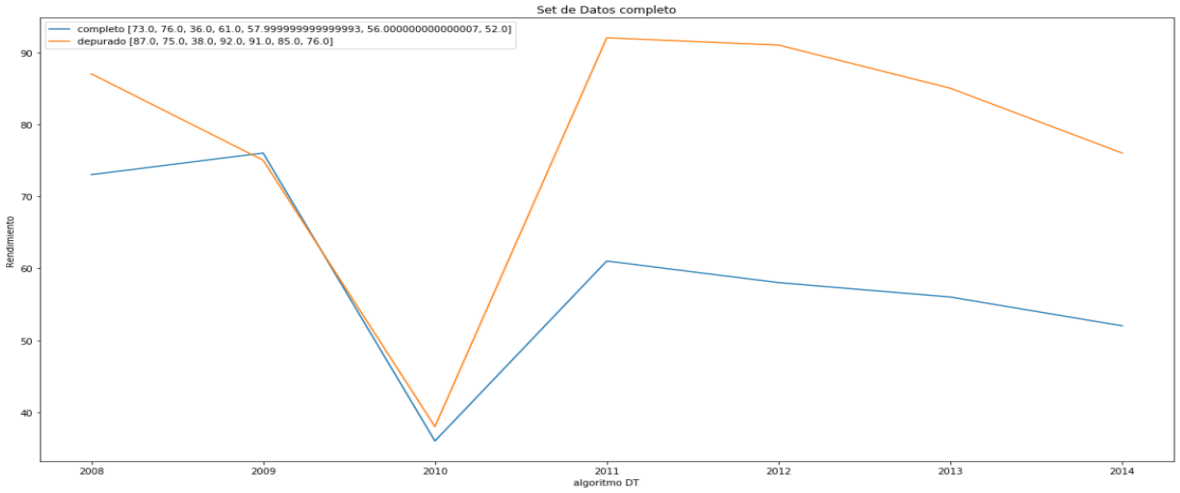


Ilustración 93. Eliminación Experimental RF SEGURIDAD SOCIAL.

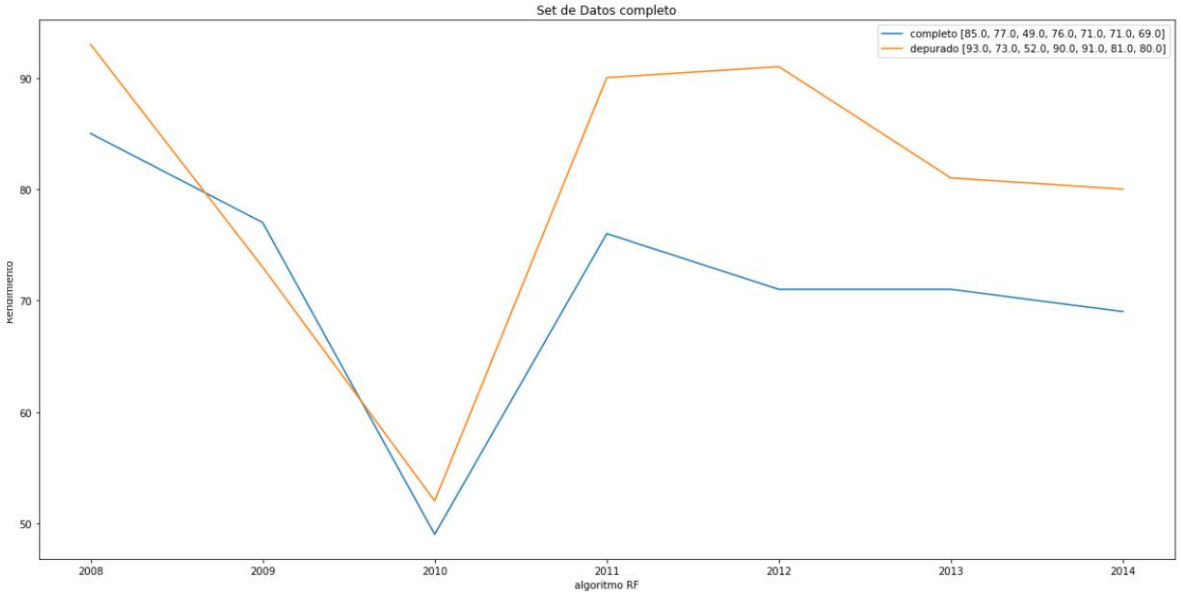


Tabla 43. Comparación anual rendimiento algoritmos seleccionados sin clase SEGURIDAD SOCIAL datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	73	76	36	61	58	56	52
RF	85	77	49	76	71	71	69

Tabla 44. Comparación anual rendimiento algoritmos seleccionados sin clase SEGURIDAD SOCIAL datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	87	75	38	92	91	85	76
RF	93	73	52	90	91	81	80

Con los gráficos anteriores se confirma que las dos clases que afectan en mayor proporción la variabilidad al modelo en su nivel, son COMUNA y ESTRATO, tal como lo proponen los promedios de los algoritmos de identificación de características, usando la eliminación de una sola variable y su reingreso posterior al eliminar una próxima variable.

ELIMINACIÓN MÉTODO BACKWARD

Veamos cómo se comporta el modelo utilizando el método original backward, las ilustraciones entre la 95 y la 107 permiten comprender el nivel de afectación de las variables al usar este método.

Ilustración 94. Método Backward RF ETNIA.

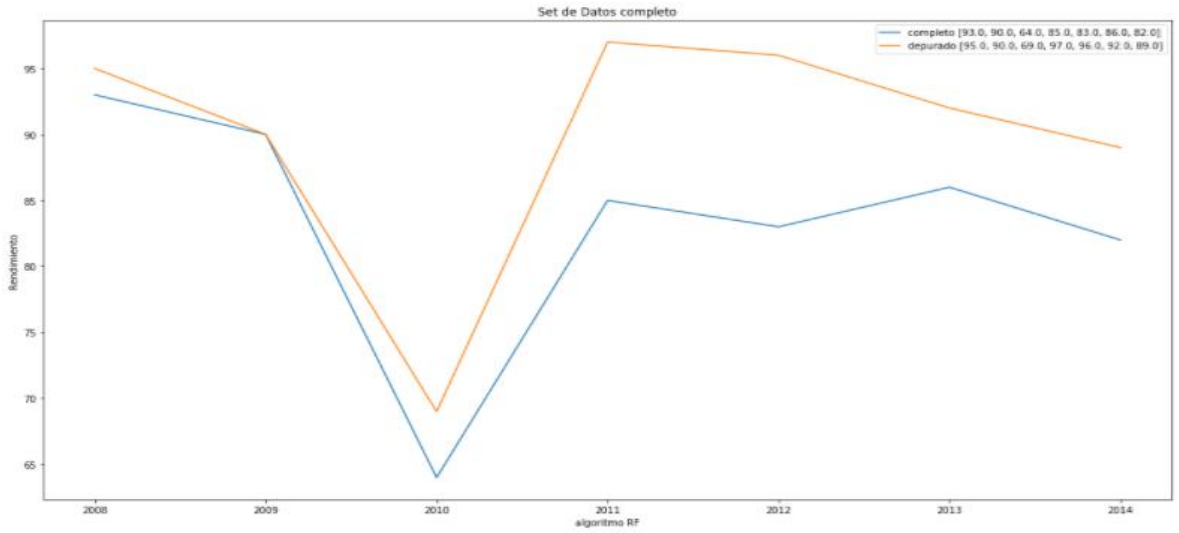


Ilustración 95. Método Backward DT ETNIA.

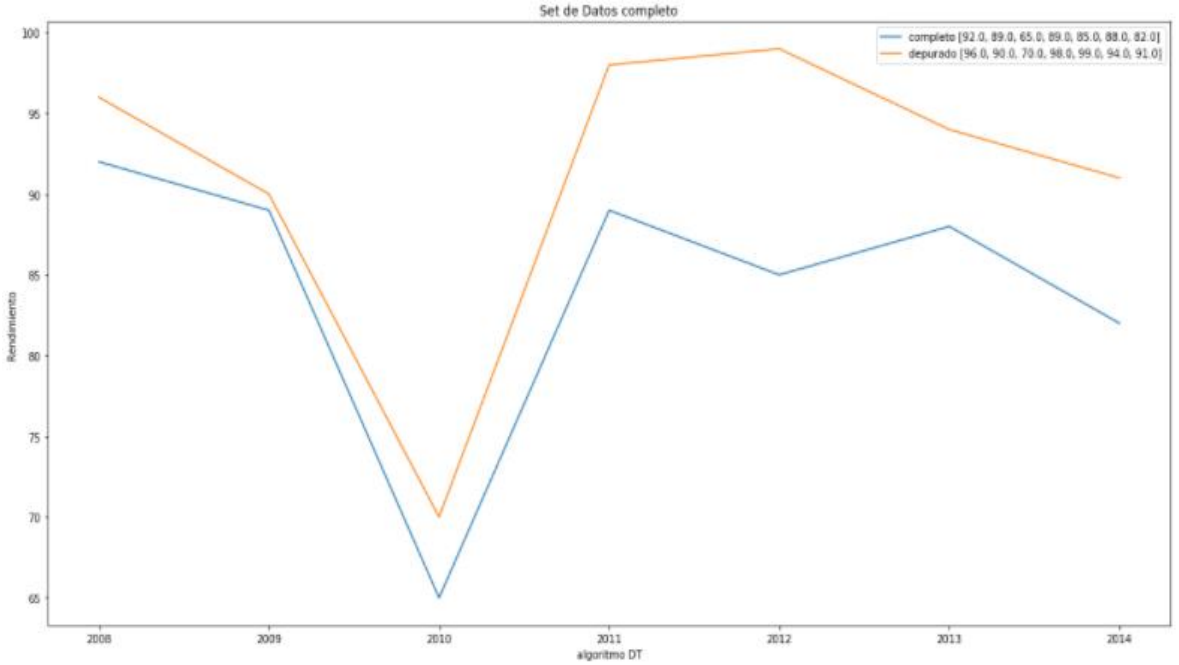


Tabla 45. Comparación anual rendimiento algoritmos seleccionados sin clase ETNIA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	92	89	65	89	85	88	82
RF	93	90	64	85	83	86	82

Tabla 46. Comparación anual rendimiento algoritmos seleccionados sin clase ETNIA datos confrmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	96	90	70	98	99	94	91
RF	95	90	69	97	96	92	89

Ilustración 96. Método Backward DT ETNIA – ÁREA.

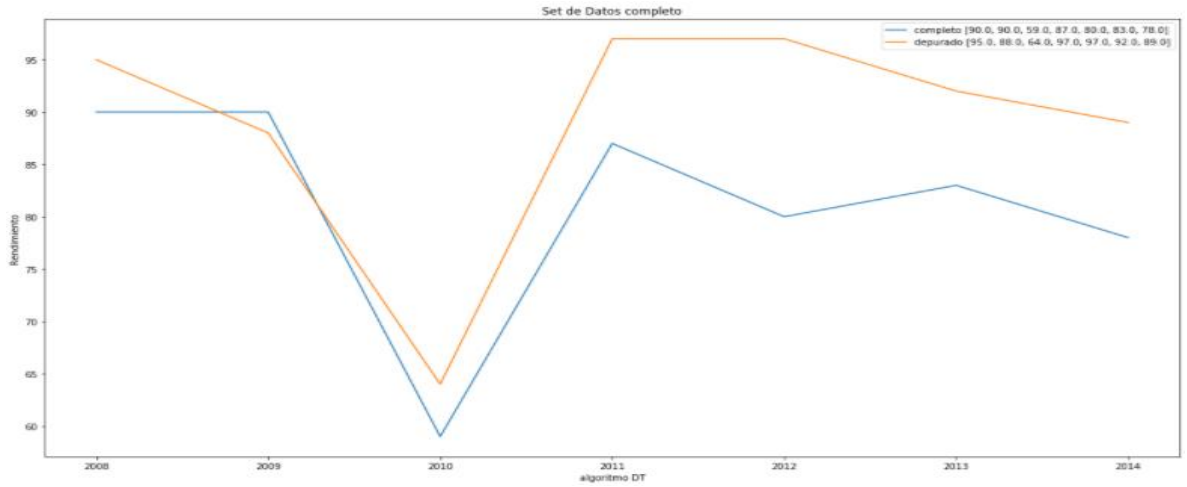


Ilustración 97. Método Backward RF ETNIA – ÁREA.

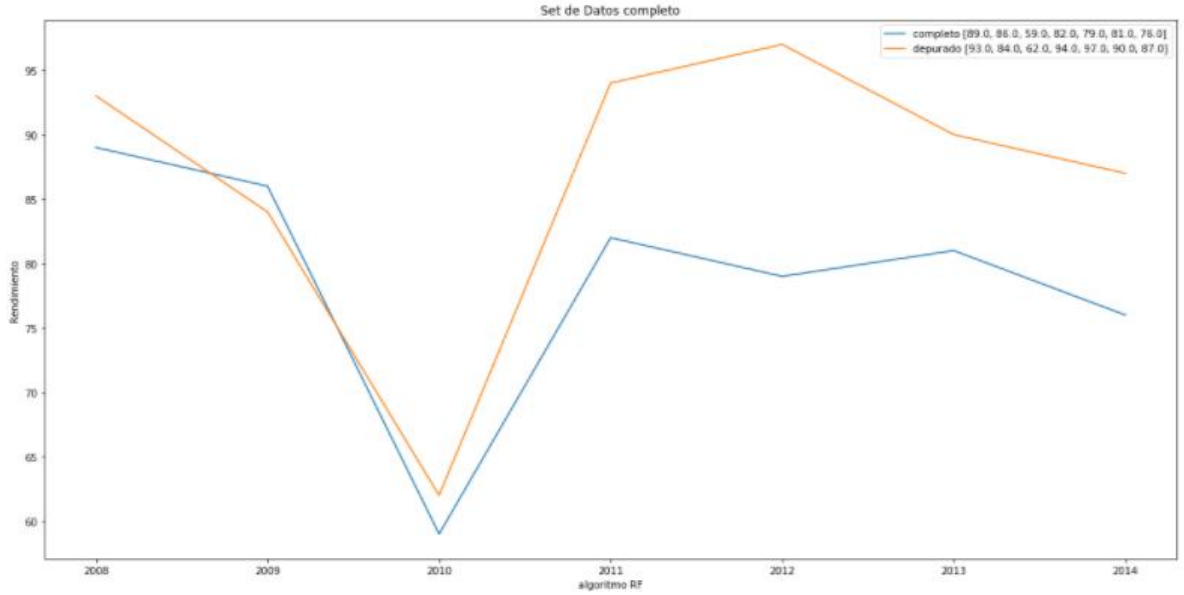


Tabla 47. Método Backward sin clase ETNIA – ÁREA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	90	90	59	87	80	83	78
RF	89	86	59	82	79	81	76

Tabla 48. Método Backward sin clase ETNIA – ÁREA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	95	88	64	97	97	92	89
RF	93	84	62	94	97	90	87

Ilustración 98. Método Backward RF ETNIA - ÁREA - ETARIO.

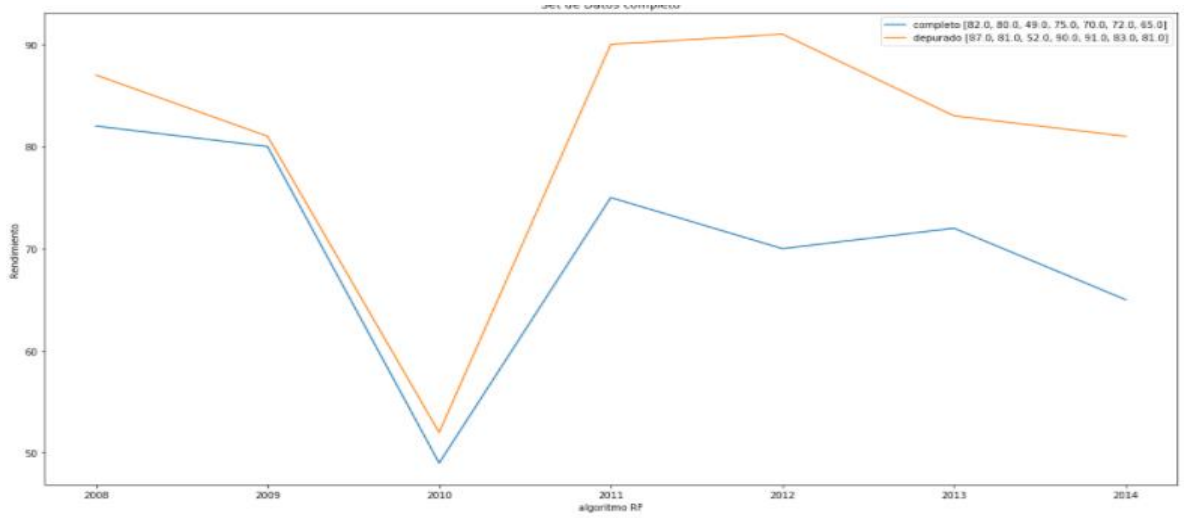


Ilustración 99. Método Backward DT ETNIA - ÁREA – ETARIO.

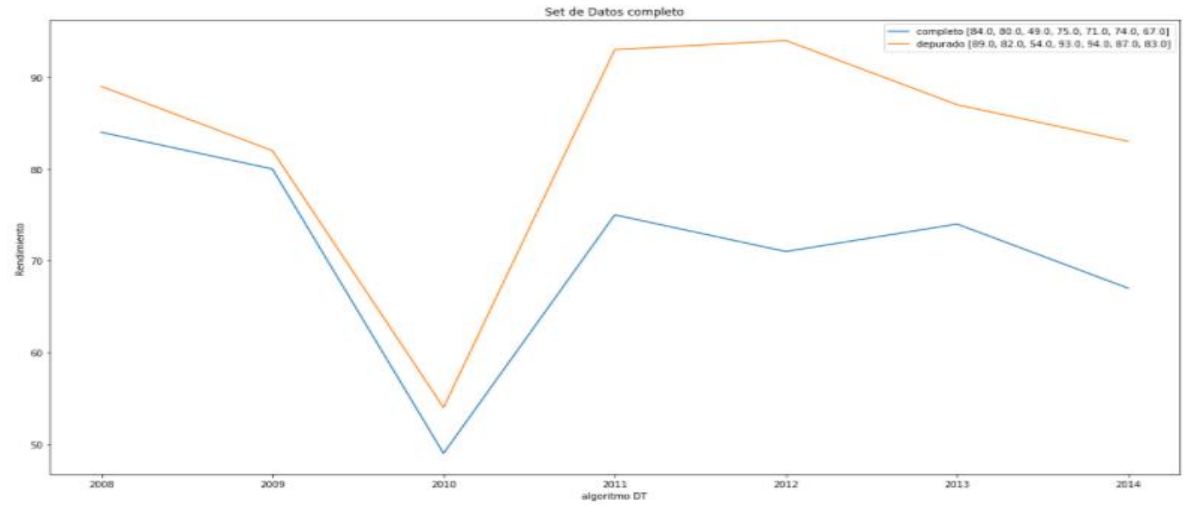


Tabla 49. Método Backward sin clase ETNIA - ÁREA – ETARIO datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	84	80	49	75	71	74	67
RF	82	80	49	75	70	72	65

Tabla 50. Método Backward sin clase ETNIA - ÁREA – ETARIO datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	89	82	54	93	94	87	83
RF	87	81	52	90	91	83	81

Ilustración 100. Método Backward DT ETNIA - ÁREA - ETARIO - SS.

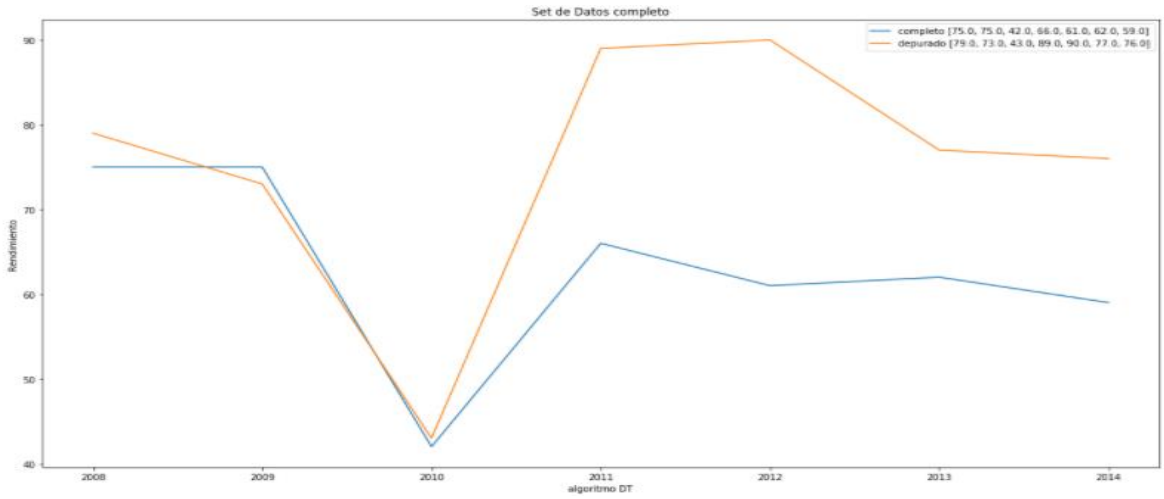


Ilustración 101. Método Backward RF ETNIA - ÁREA - ETARIO - SS.

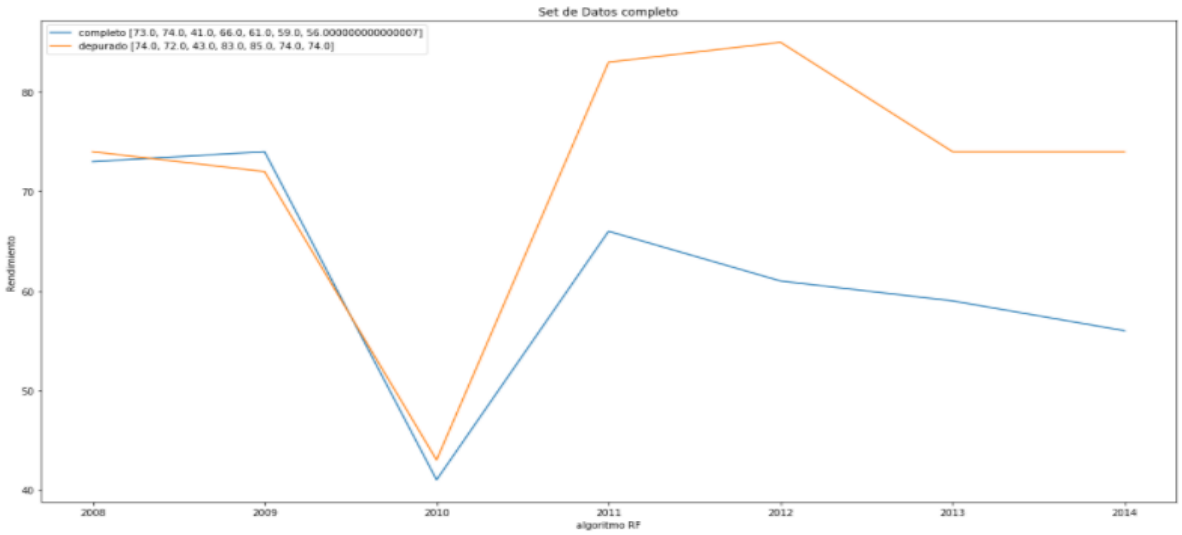


Tabla 51. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	75	75	42	66	61	62	59
RF	73	74	41	66	61	59	56

Tabla 52. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	79	73	43	89	90	77	76
RF	74	72	43	83	85	74	74

Ilustración 102. Método Backward RF ETNIA - ÁREA - ETARIO - SS - OCUPACIÓN.

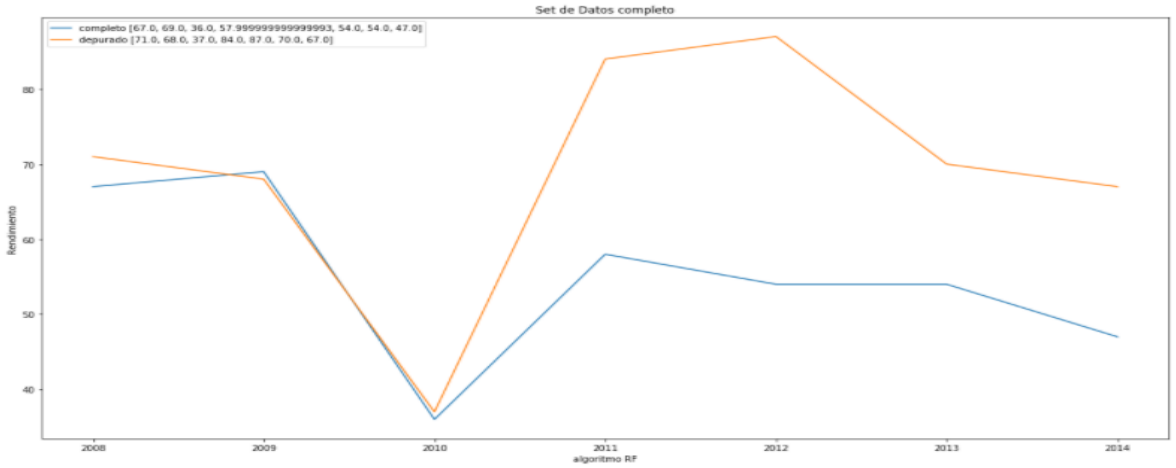


Ilustración 103. Método Backward DT ETNIA - ÁREA - ETARIO - SS - OCUPACIÓN.

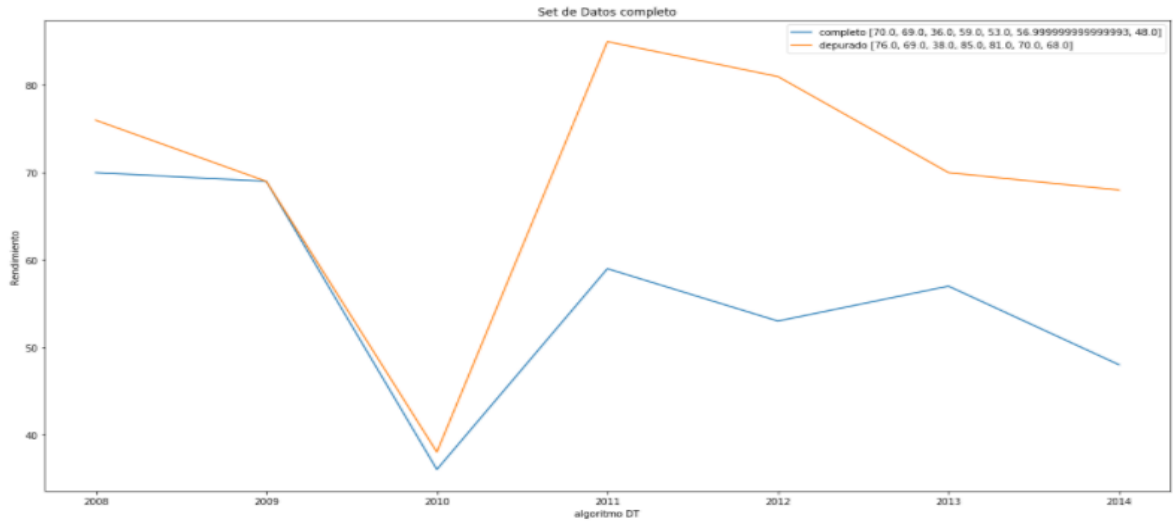


Tabla 53. Método Backward sin clase ETNIA - ÁREA - ETARIO - SS – OCUPACIÓN datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	70	69	36	59	53	57	48
RF	67	69	36	58	54	54	47

Tabla 54. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS – OCUPACIÓN datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	76	69	38	85	81	70	68
RF	71	68	37	84	87	70	67

Ilustración 104. Método Backward RF ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO.

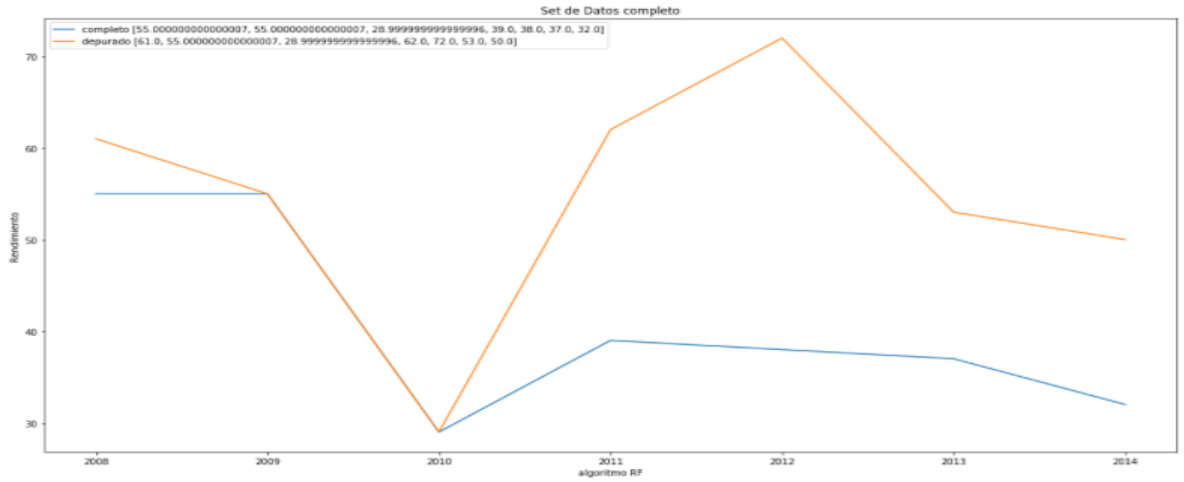


Ilustración 105. Método Backward DT ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO.

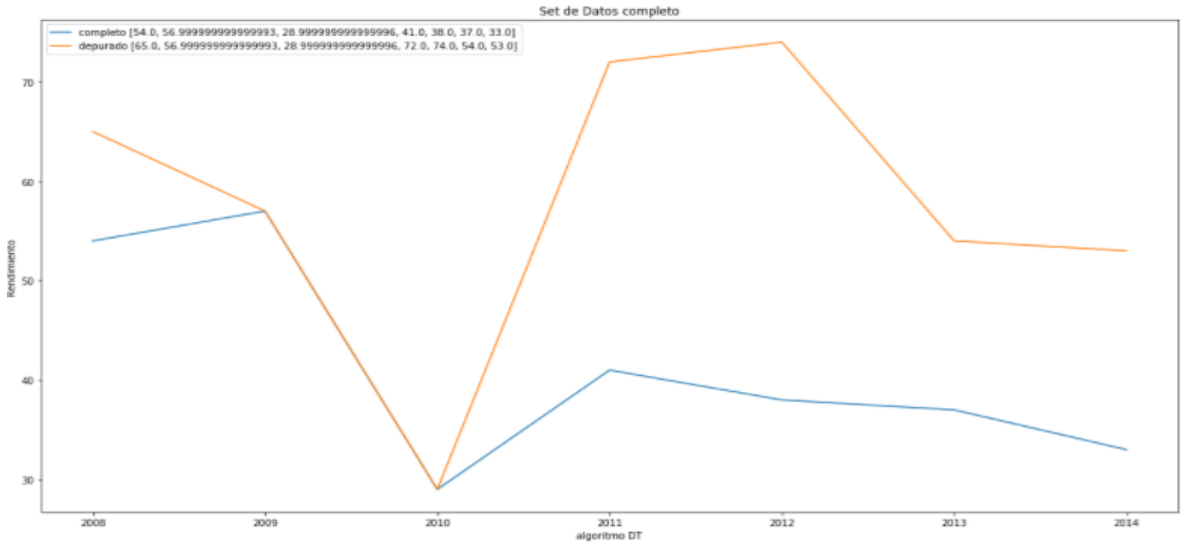


Tabla 55. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN – ESTRATO datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	54	57	29	41	38	37	33
RF	55	55	29	39	38	37	32

Tabla 56. Método Backward sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN – ESTRATO datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	65	57	29	72	74	54	53
RF	61	55	29	62	72	53	50

Ilustración 106. Método Mayor-Menor RF ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN – ESTRATO - COMUNA.

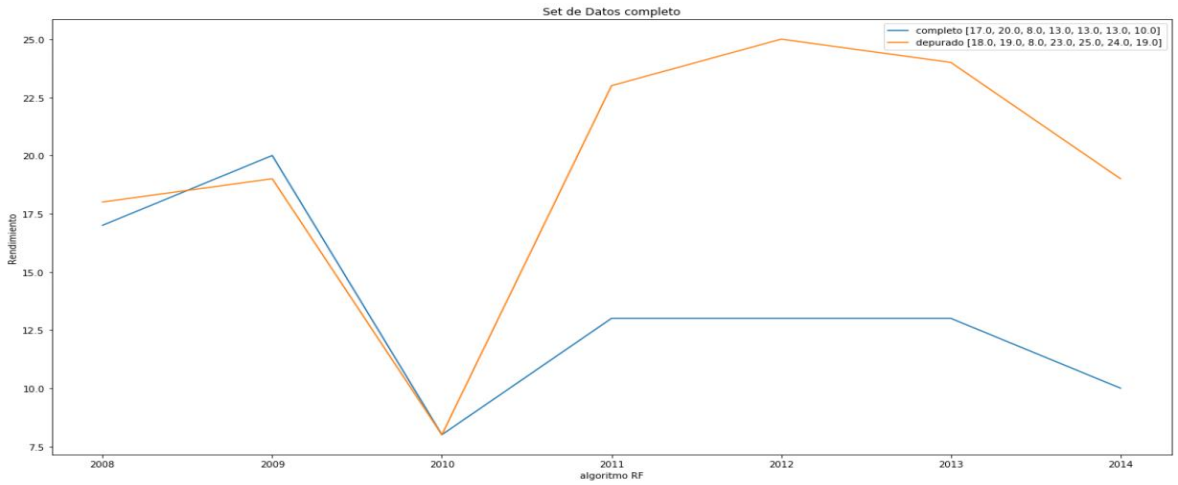


Ilustración 107. Método Mayor-Menor DT ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO - COMUNA.

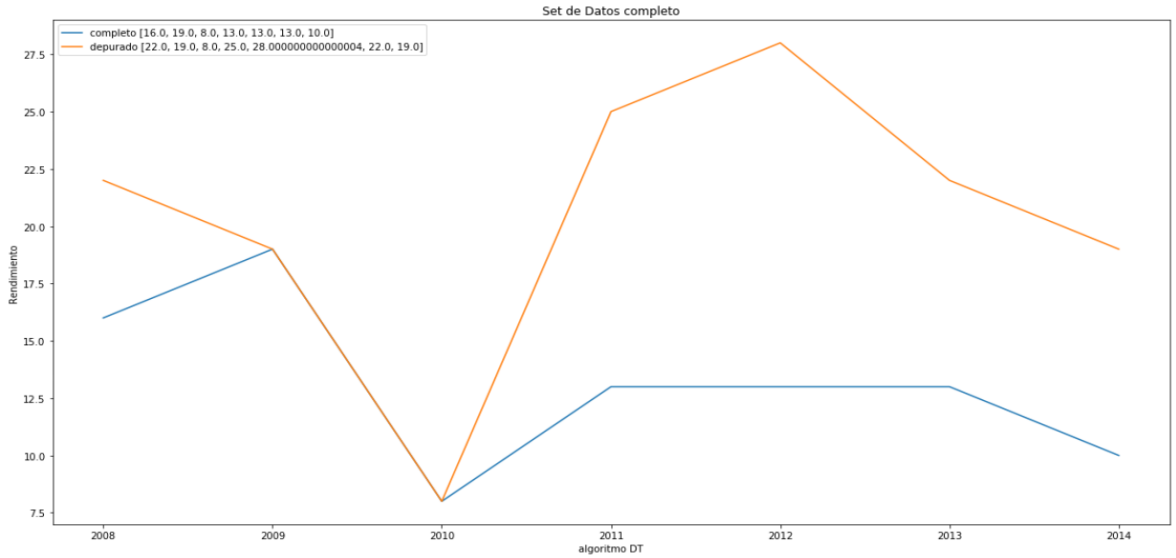


Tabla 57. Método Mayor-Menor sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO – COMUNA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	16	19	8	13	13	13	10
RF	17	20	8	13	13	13	10

Tabla 58. Método Mayor-Menor sin clase ETNIA - ÁREA - ETARIO – SS - OCUPACIÓN - ESTRATO – COMUNA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	22	19	8	25	28	22	19
RF	18	19	8	23	25	24	19

Con los gráficos y las tablas obtenidas con el método backward, se puede entender la manera en que cada una de las clases afecta la variabilidad de la muestra, y el grado de importancia que representa cada una de las clases en el modelo y que afecta directamente la efectividad en la predicción de cada clasificador. Es claro que si existe una relación espacial en los casos de las personas que contraen el virus del dengue y se evidencia por las comunas donde habitan, también se puede decir que la parte económica es identificable ya que los estratos son diferenciales para la muestra y aportan un importante valor a la interpretación del modelo, en la parte social con la ocupación y el tipo de seguridad social también son factores importantes para la creación y explicación de la variabilidad del modelo, entonces bajo esta óptica si es posible caracterizar la influencia de las variables sociales, económicas y espaciales en enfermedades trasmitidas por vectores usando técnicas de machine learning. Es importante tener en cuenta que el uso de técnicas de evaluación de variables explicativas para describir un modelo son de gran importancia para lograr muy buenos resultados en los análisis de datos (Shelley Derksen, 1992), el apoyo de estudios que implementan técnicas para evaluar variables son fundamentales en el estudio de nuevos modelos,

aunque sean análisis en campos diferentes al de este trabajo como por ejemplo conocer cuales clases explican mejor un fenómeno como la detección de actividades cerebrales que realizan en la publicación de la revista fronteras de la neurociencia (Marino Andres. Alvarez-Meza, 2017) y conocer de manera alternativa como aplicar este tipo de técnicas a set de datos pequeños y obtener el mejor provecho de estos (Thompson, 1995).

ELIMINACIÓN MÉTODO MAYOR IMPORTANCIA A MENOR IMPORTANCIA

Contrario a lo que plantea el método backward, se utilizará esta propuesta de eliminar la clase más importante hasta la menos importante del modelo, para observar quá tan rápido decae la viabilidad del modelo para cada uno de los años de ambos sets de datos, esto se puede observar desde la ilustración 108 hasta la ilustración 121.

Ilustración 108. Método Mayor-Menor DT COMUNA.

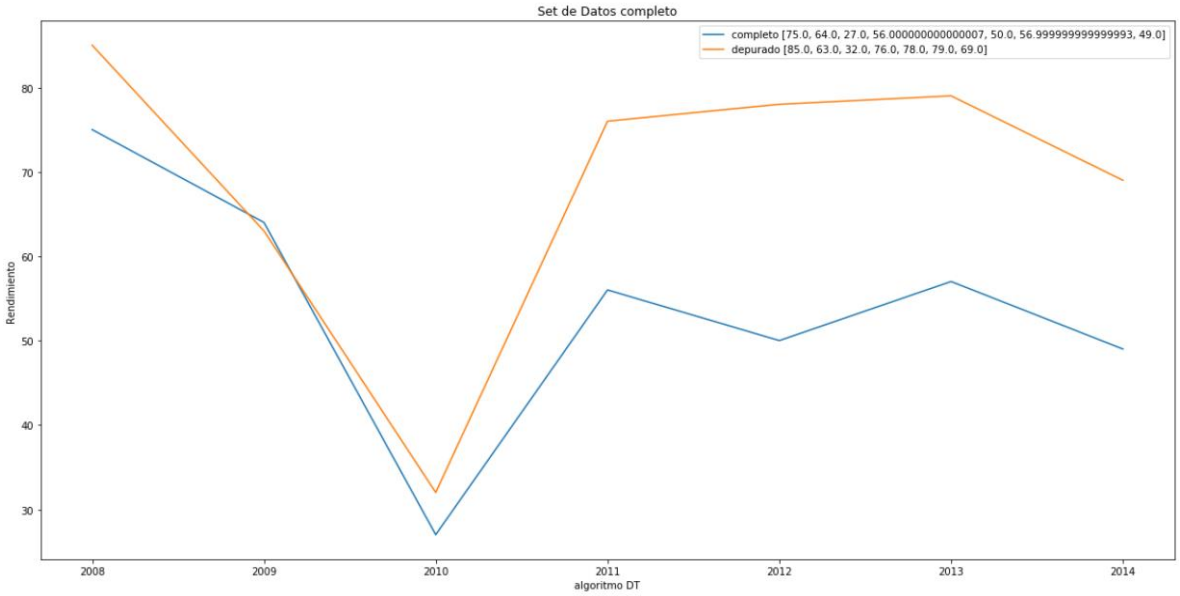


Ilustración 109. Método Mayor-Menor RF COMUNA.

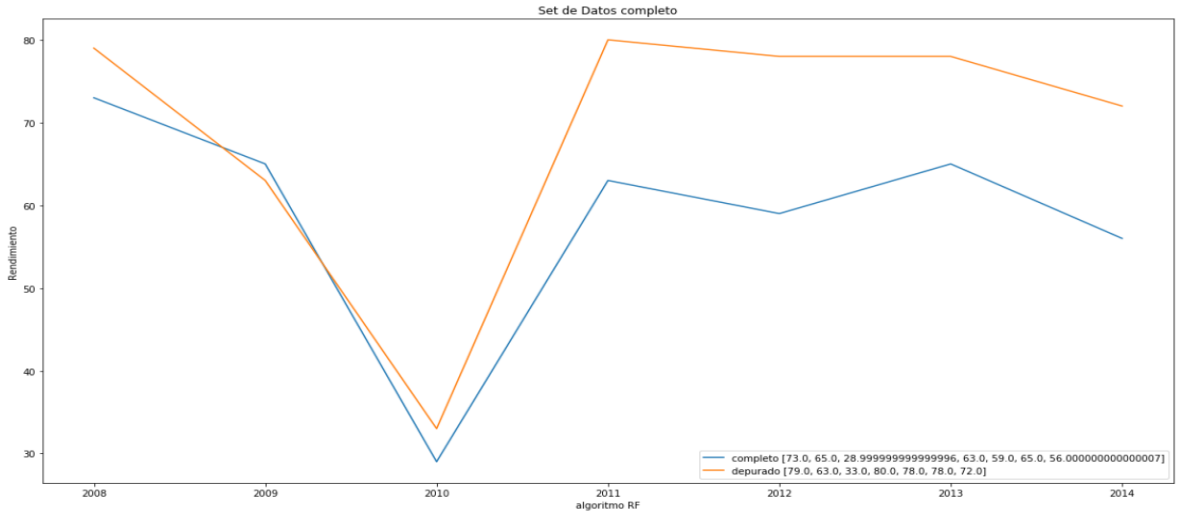


Tabla 59. Método Mayor-Menor sin clase COMUNA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	75	64	27	56	50	57	49
RF	73	65	29	63	59	65	56

Tabla 60. Método Mayor-Menor sin clase COMUNA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	85	63	32	76	78	79	69
RF	79	63	33	80	78	78	72

Ilustración 110. Método Mayor-Menor RF COMUNA - ESTRATO.

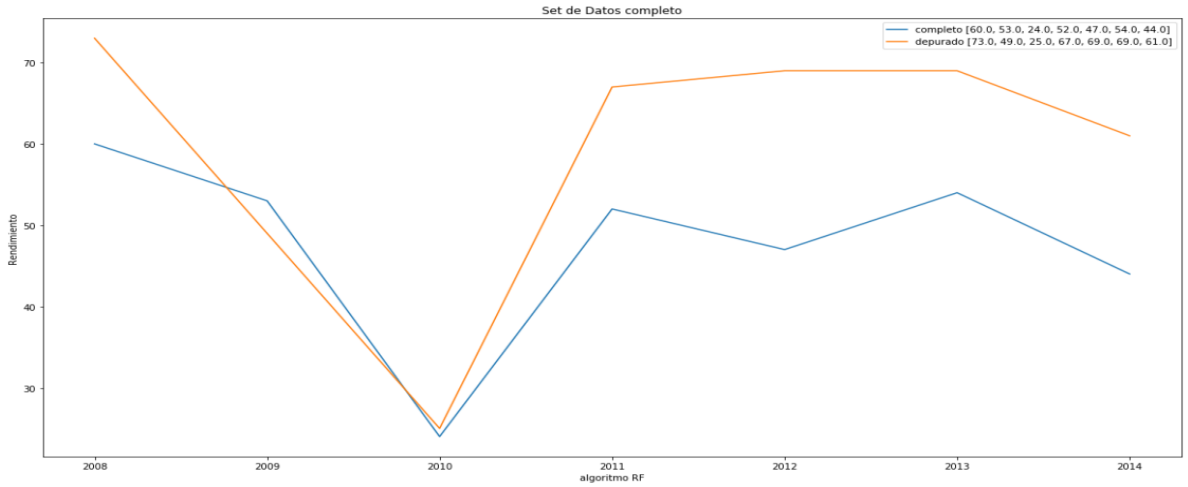


Ilustración 111. Método Mayor-Menor DT COMUNA - ESTRATO.

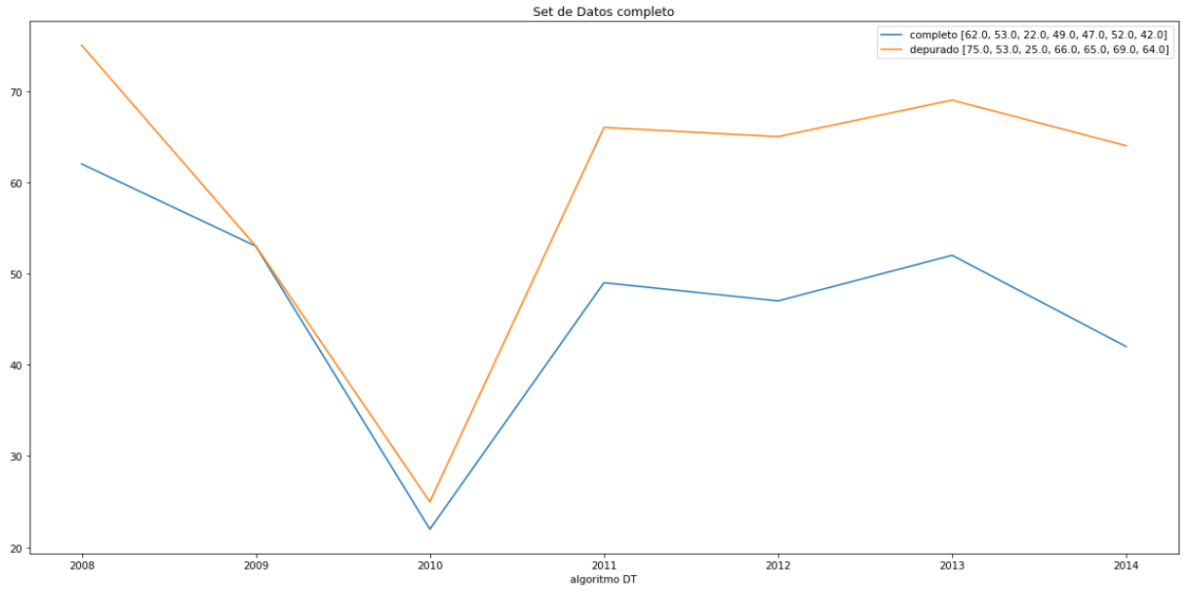


Tabla 61. Método Mayor-Menor sin clase COMUNA - ESTRATO datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	62	53	22	49	47	52	42
RF	60	53	24	52	47	54	44

Tabla 62. Método Mayor-Menor sin clase COMUNA - ESTRATO datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	75	53	25	66	65	69	64
RF	73	49	25	67	69	69	61

Ilustración 112. Método Mayor-Menor RF COMUNA - ESTRATO - OCUPACIÓN.

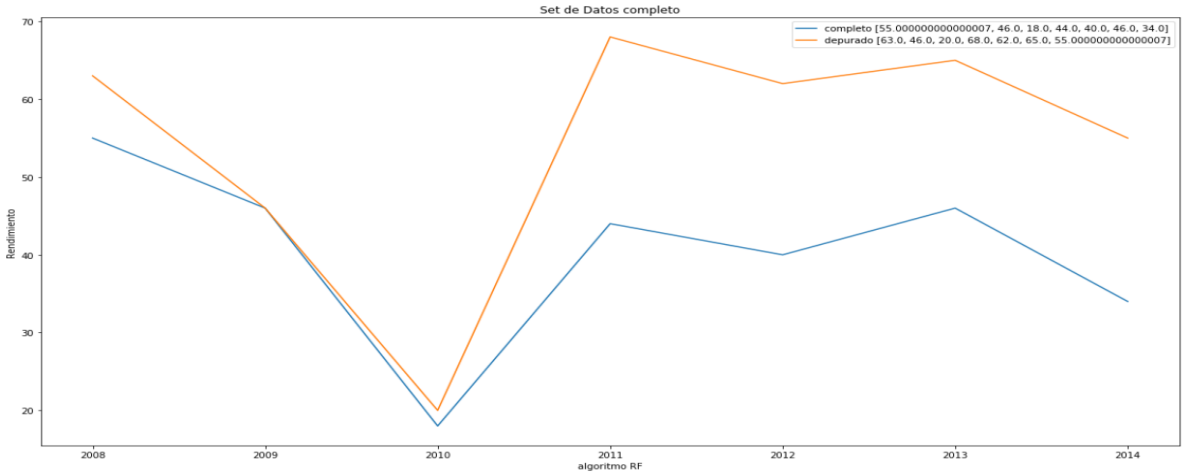


Ilustración 113. Método Mayor-Menor DT COMUNA - ESTRATO - OCUPACIÓN.

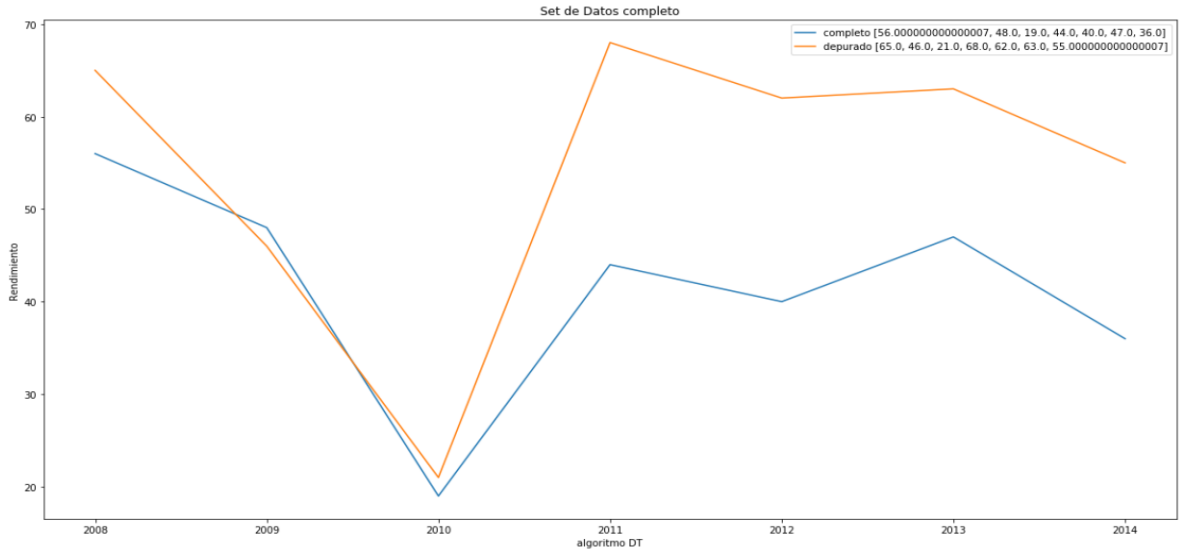


Tabla 63. Método Mayor-Menor sin clase COMUNA – ESTRATO - OCUPACIÓN datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	54	48	19	44	40	47	36
RF	55	46	18	44	40	46	34

Tabla 64. Método Mayor-Menor sin clase COMUNA - ESTRATO - OCUPACIÓN datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	65	46	21	68	62	63	55
RF	63	46	20	68	62	65	55

Ilustración 114. Método Mayor-Menor DT COMUNA – ESTRATO – OCUPACIÓN - SS.

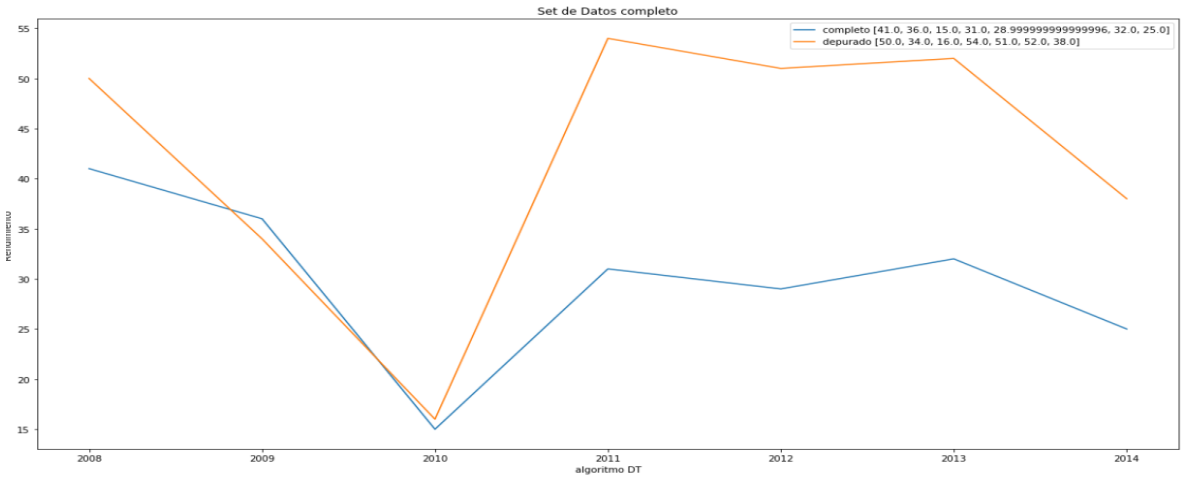


Ilustración 115. Método Mayor-Menor RF COMUNA – ESTRATO – OCUPACIÓN - SS.

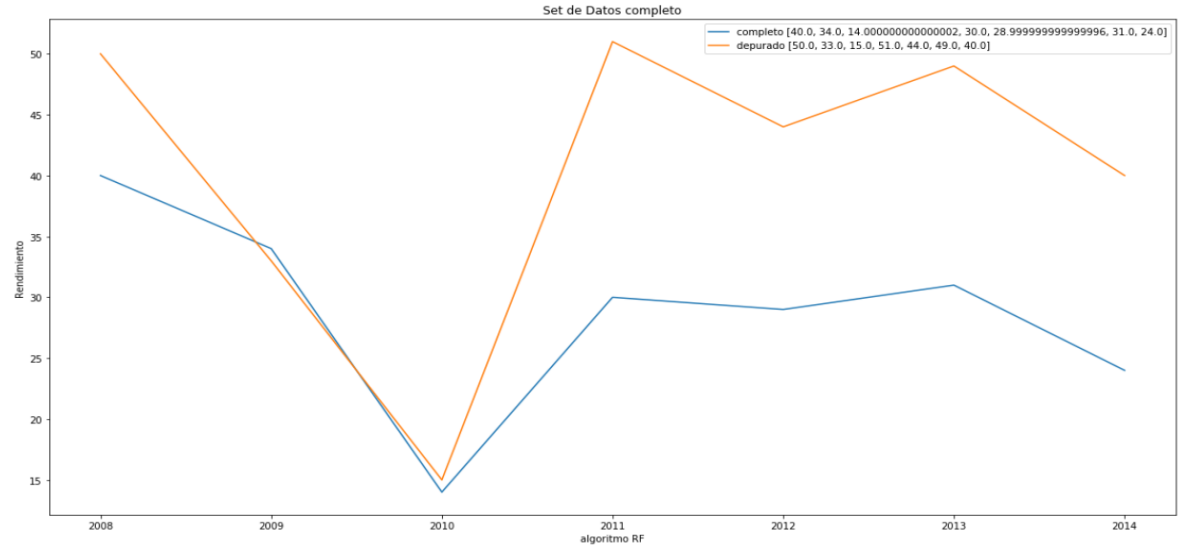


Tabla 65. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN - SS datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	41	36	15	31	29	32	25
RF	40	34	14	30	29	31	24

Tabla 66. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN - SS datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	50	34	16	54	51	52	38
RF	50	33	15	51	44	49	40

Ilustración 116. Método Mayor-Menor RF COMUNA – ESTRATO – OCUPACIÓN – SS - ETARIO.

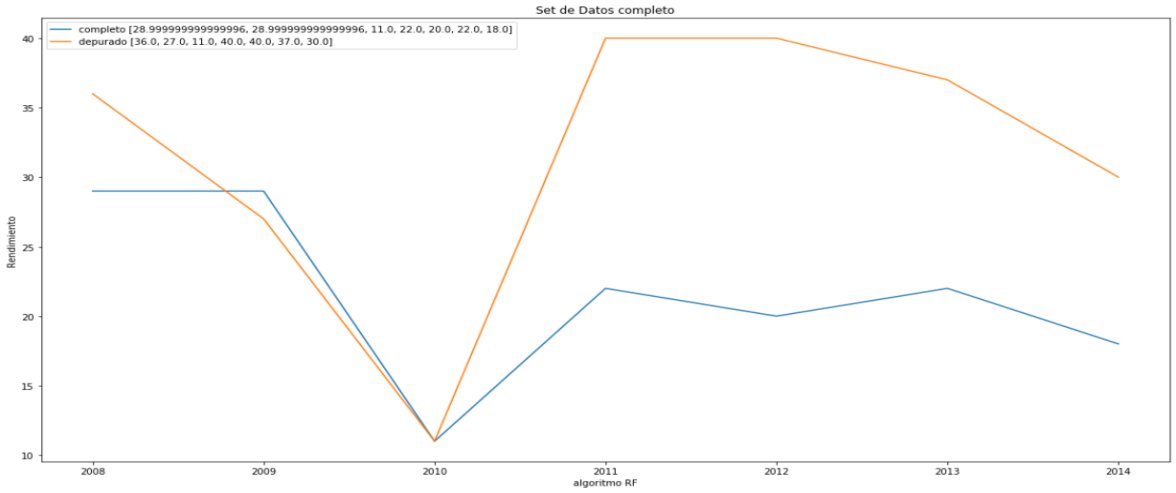


Ilustración 117. Método Mayor-Menor DT COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO

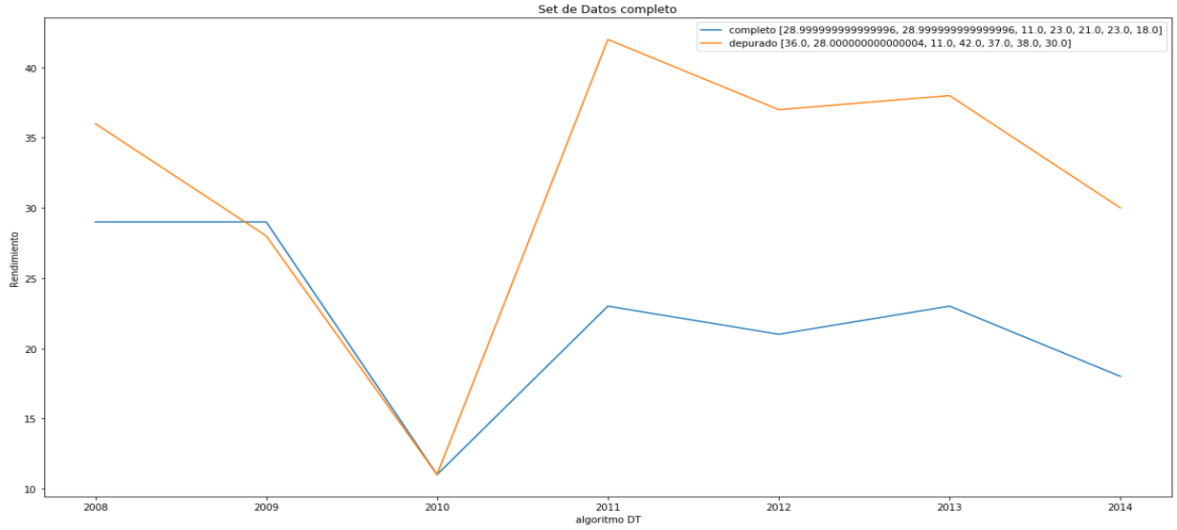


Tabla 67. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN – SS - ETARIO datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	29	29	11	23	21	23	18
RF	29	29	11	22	20	22	18

Tabla 68. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN – SS - ETARIO datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	36	28	11	42	37	38	30
RF	36	27	11	40	40	37	30

Ilustración 118. Método Mayor-Menor DT COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA.

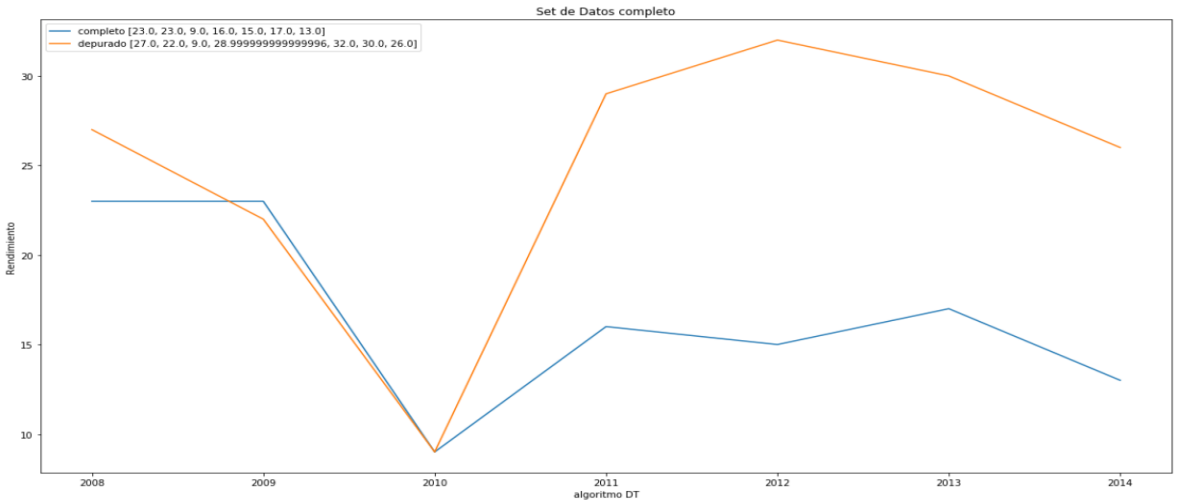


Ilustración 119. Método Mayor-Menor RF COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA.

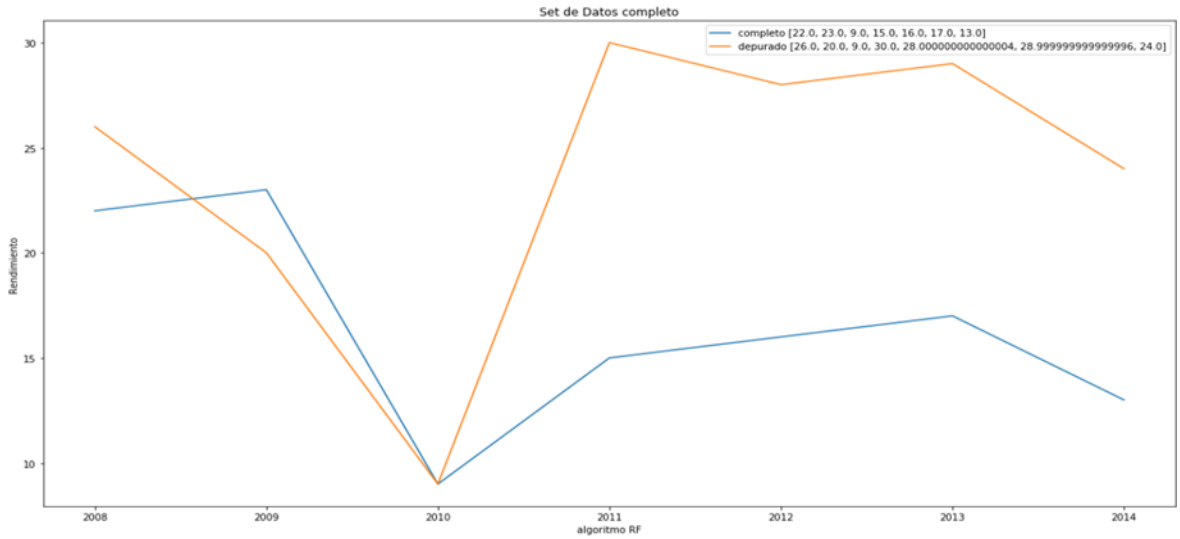


Tabla 69. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO - ÁREA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	23	23	9	16	15	17	13
RF	22	23	9	15	16	17	13

Tabla 70. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO - ÁREA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	27	22	9	29	32	30	26
RF	26	20	9	30	28	29	24

Ilustración 120. Método Mayor-Menor RF COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA – ETNIA.

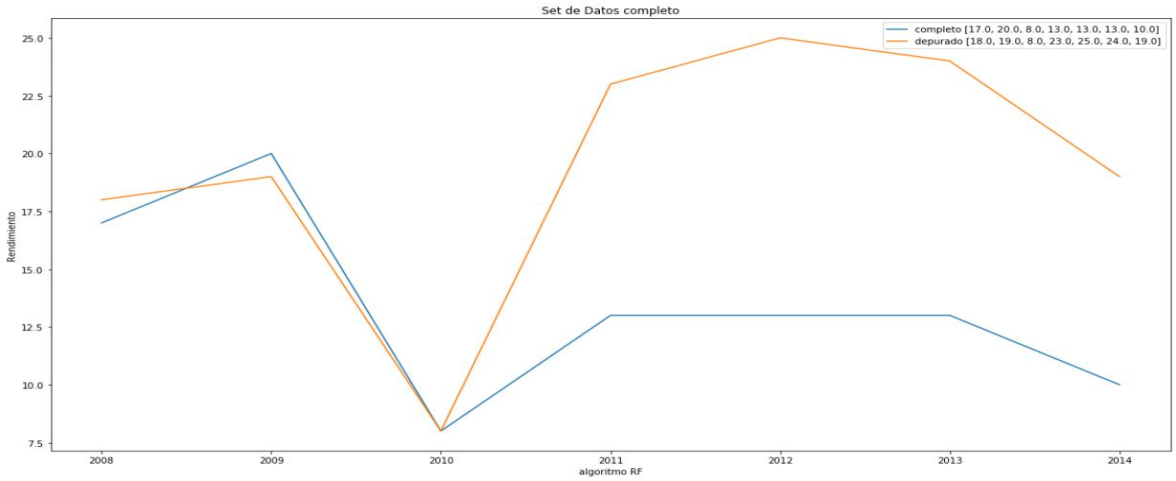


Ilustración 121. Método Mayor-Menor DT COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA - ETNIA.

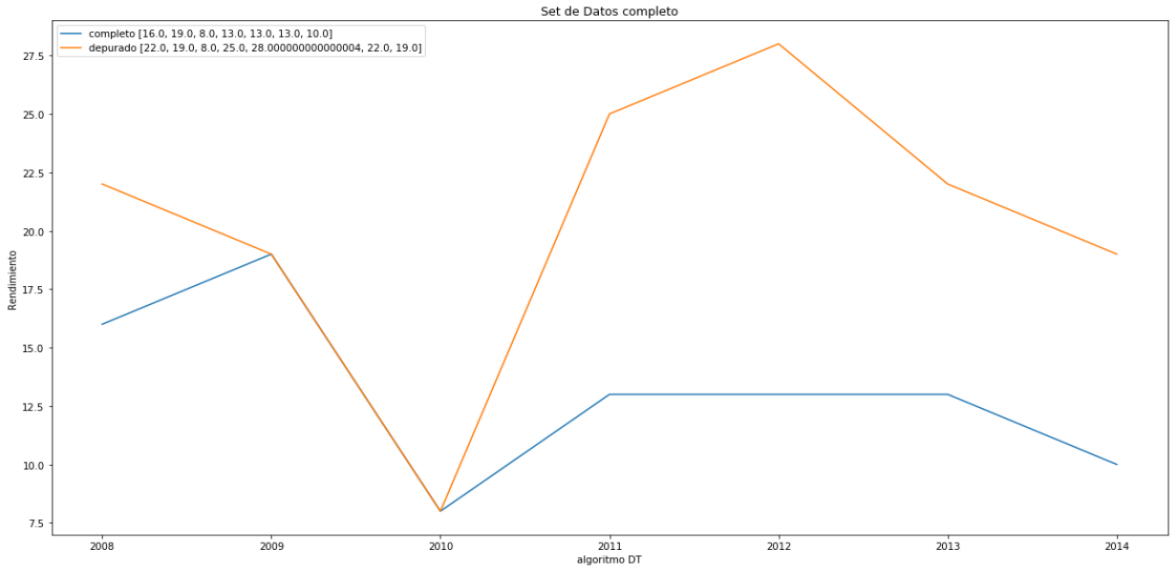


Tabla 71. Método Mayor-Menor sin clase COMUNA – ESTRATO – OCUPACIÓN – SS – ETARIO – ÁREA - ETNIA datos completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	16	19	8	13	13	13	10
RF	17	20	8	13	13	13	10

Tabla 72. Método Mayor-Menor sin clase COMUNA - ESTRATO – OCUPACIÓN – SS – ETARIO - ÁREA - ETNIA datos confirmados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	22	19	8	25	28	22	19
RF	18	19	8	23	25	24	19

En esta propuesta de eliminación de la variable más importante a la de menor importancia también se observa que al eliminar las dos variables más importantes del set de datos como son comuna y estrato, el modelo decae de manera muy fuerte de un máximo de 40% sobre la explicación de la variabilidad del modelo, en esta eliminación se aprecia rápidamente el impacto de la falta de esas clases comparado con el método backward que hasta casi el final se pudo evidenciar que el modelo era muy poco viable en cada uno de los casos anuales.

X. ZONAS A EVALUAR EN EL SET DE DATOS

El número de casos en cada municipio del departamento de Risaralda son el recurso más importante del set de datos, pero es de recordar como se vió en ilustraciones anteriores algunos municipios reportaban cifras anuales muy bajas de hasta algo superior a sesenta casos anuales, y en algunos casos anualmente reportaron cero casos, teniendo como distinción el 2010 donde fue anormal el fenómeno y todos los municipios reportaron casos, para evitar sobre entrenamiento del algoritmo sobre la muestra se tomarán en cuenta los municipios de Risaralda que reporten sin falla casos anualmente y que por lo menos supere la cifra anual mínima de 5 casos, por esta razón se vuelve a depurar la información y quedan cuatro municipios para esto, que son: Pereira, Dosquebradas, Santa Rosa de Cabal y La Virginia, es un cambio representativo si se tiene en cuenta que los estos cuatro municipios representan el 89.8% del total del set de datos y el 91% de total de datos depurados.

XI. NIVEL DE PREDICCIÓN DE LOS CASOS DE DENGUE

Como el objetivo del machine learning se basa en predecir fenómenos a partir de la capacidad de aprendizaje del conjunto de datos que se usaron para entrenar, se aprecia cómo se comportan los clasificadores, tanto para la muestra con todos los municipios como para el caso particular de los municipios elegidos para estudio focalizado específicamente en Pereira, Dosquebradas, Santa Rosa y La Virginia. Al observar su comportamiento mediante las ilustraciones 122 y 123.

Ilustración 122. Capacidad de Aprendizaje DT Todos los datos.

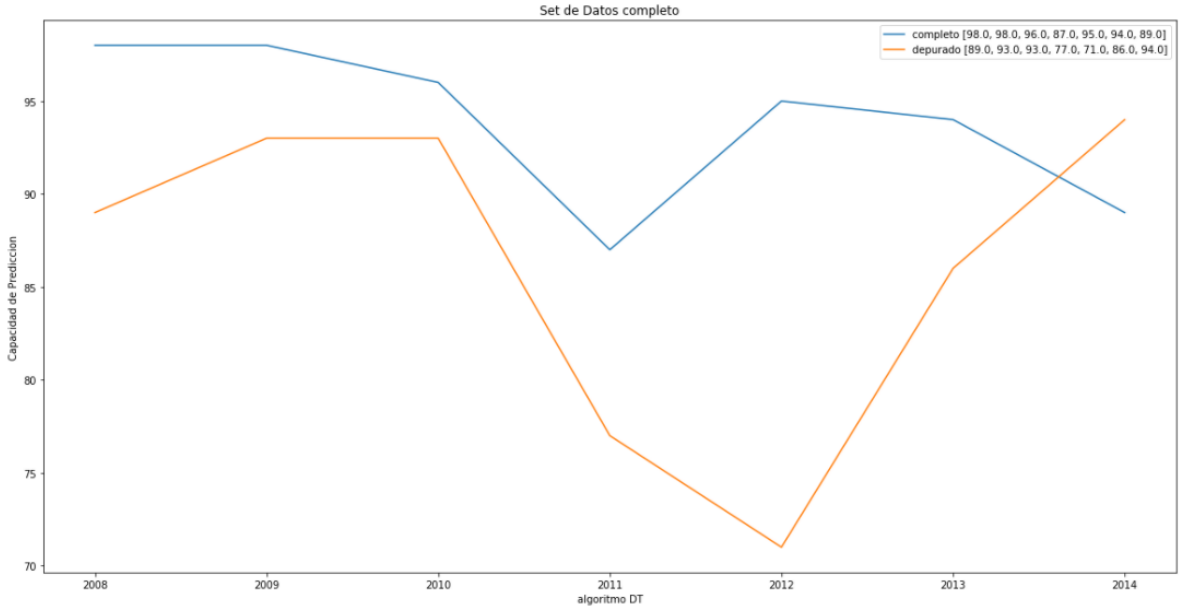


Ilustración 123. Capacidad de Aprendizaje RF Todos los datos.

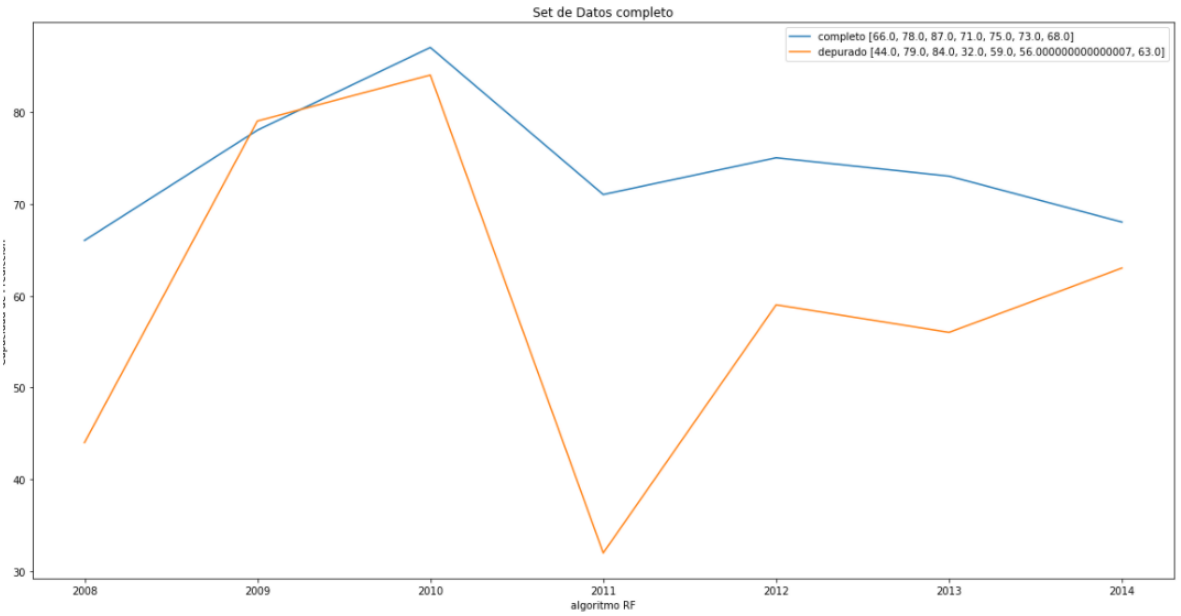


Tabla 73. Capacidad Predictiva datos Completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	98	98	96	87	95	94	89
RF	66	78	87	71	75	73	68

Tabla 74. Capacidad Predictiva datos Depurados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	89	93	93	77	71	86	94
RF	44	79	84	32	59	56	63

Realizando la evaluación de los datos arrojados por los algoritmos tipo árbol en las ilustraciones 122 y 123, y resumidos en las tablas anteriores, se percibe que la capacidad de predicción del clasificador random forest decayó notablemente y en algunos casos entrega la mitad del valor de la predicción que ofrece el árbol de decisión, existen fenómenos en las máquinas de aprendizaje en los cuales los valores que arrojan al final de todo el ciclo de machine learning son bajos o descartables, esos problemas están ligados generalmente a temas de sobreajuste u OverFitting o también por el sobre entrenamiento u OverTraining. En el OverTraining el algoritmo no logra aprender muy bien las características de los datos que usa para aprender, entonces cuando se llega a la fase de predicción esta máquina hace una clasificación mala, en el OverFitting se aprende muy bien de los datos de aprendizaje, pero al momento de la predicción de nuevos datos que no conoce, la máquina de aprendizaje no consigue clasificar de buena manera los ejemplos que no ha visto antes.

Es posible que el random forest que evalúa los datos de este trabajo esté sufriendo de OverFitting, parece ser que a menor cantidad de datos con que cuente el algoritmo para entrenar con este set de datos, menor es la capacidad de realizar una buena clasificación, esta hipótesis será puesta a prueba con la evaluación de los cuatro municipios con mayor proporción de información de casos entre las ilustraciones 124 a la 131.

Ilustración 124. Capacidad de Predicción RF Pereira

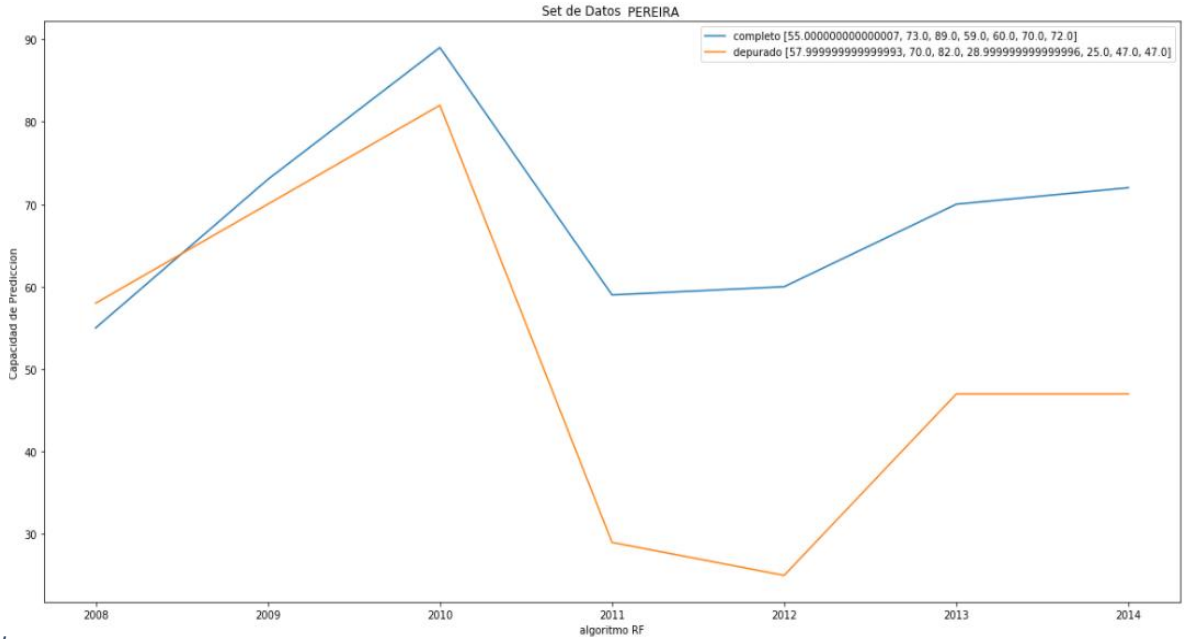


Ilustración 125. Capacidad de Predicción DT Pereira.

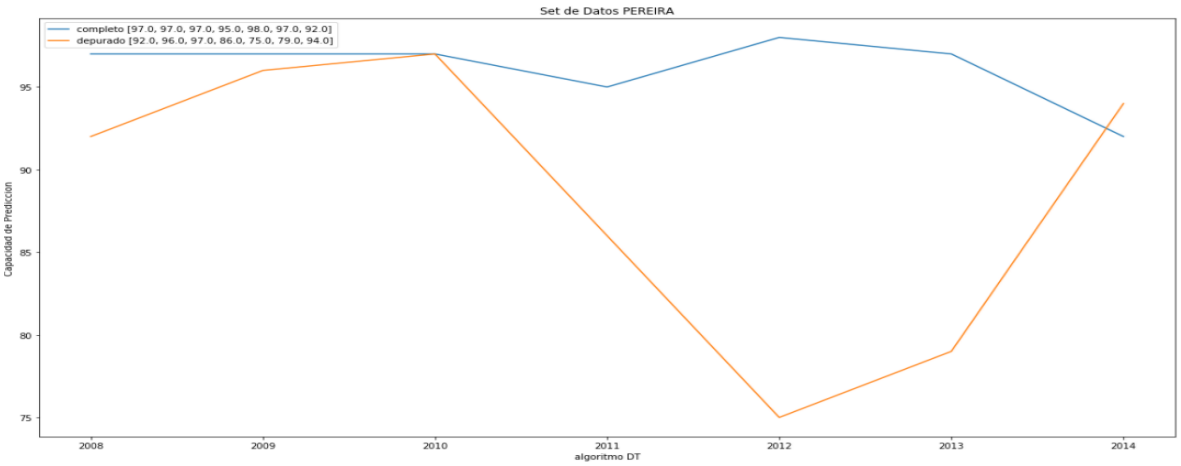


Tabla 75. Capacidad Predictiva Pereira datos Completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	97	97	97	95	98	97	92
RF	55	73	89	59	60	70	72

Tabla 76. Capacidad Predictiva Pereira datos Depurados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	92	96	97	86	75	79	94
RF	58	70	82	29	25	47	47

Ilustración 126. Capacidad de Predicción DT Dosquebradas.

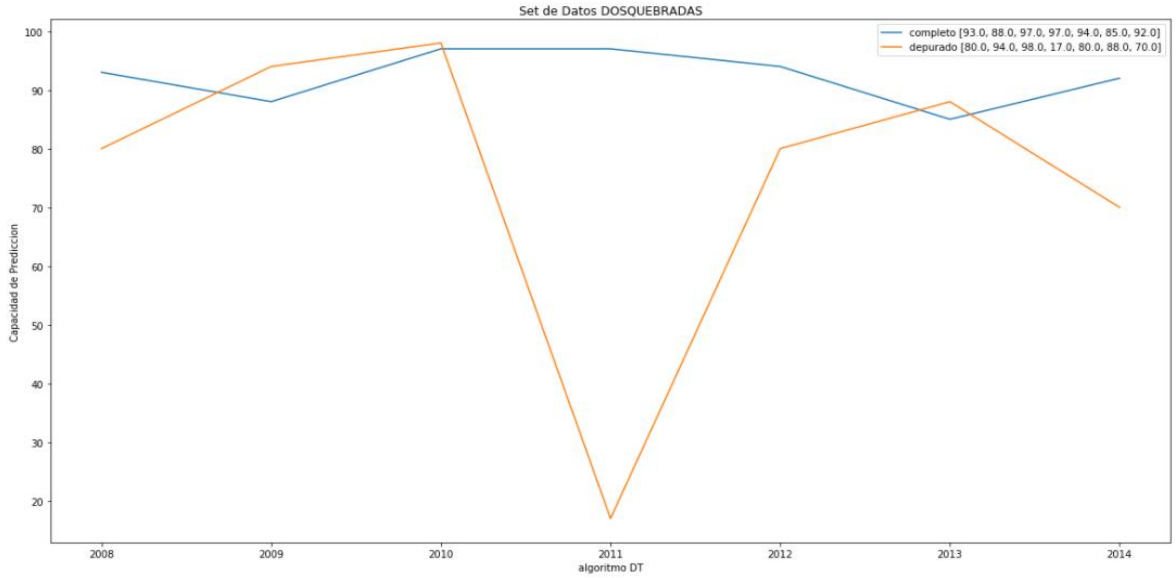


Ilustración 127. Capacidad de Predicción RF Dosquebradas.

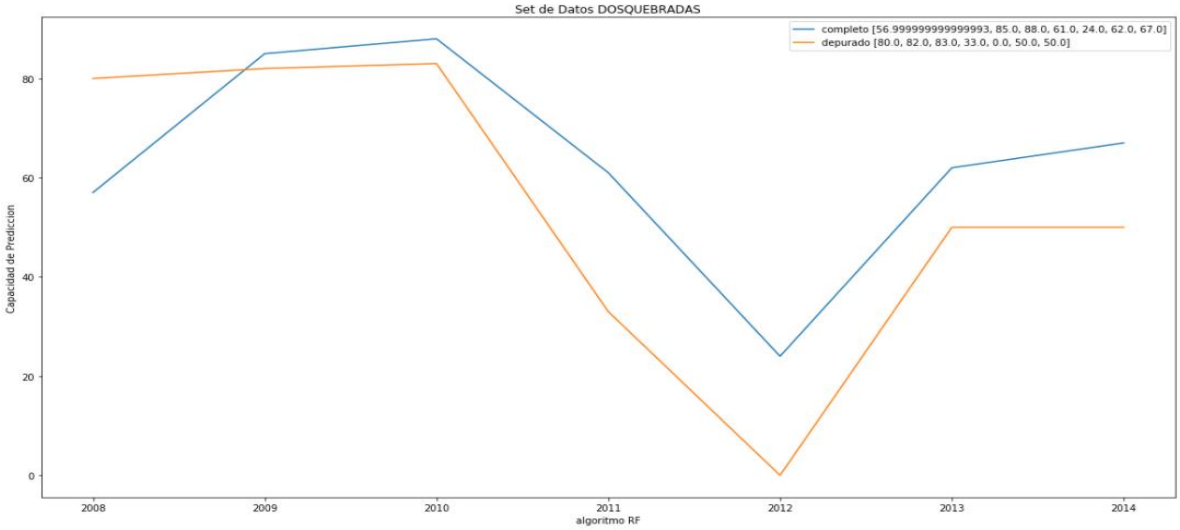


Tabla 77. Capacidad Predictiva Dosquebradas datos Completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	93	88	97	97	94	85	92
RF	57	85	88	61	24	62	67

Tabla 78. Capacidad Predictiva Dosquebradas datos Depurados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	80	94	98	17	80	88	70
RF	80	82	83	33	0	50	50

Ilustración 128. Capacidad de Predicción DT La Virginia.

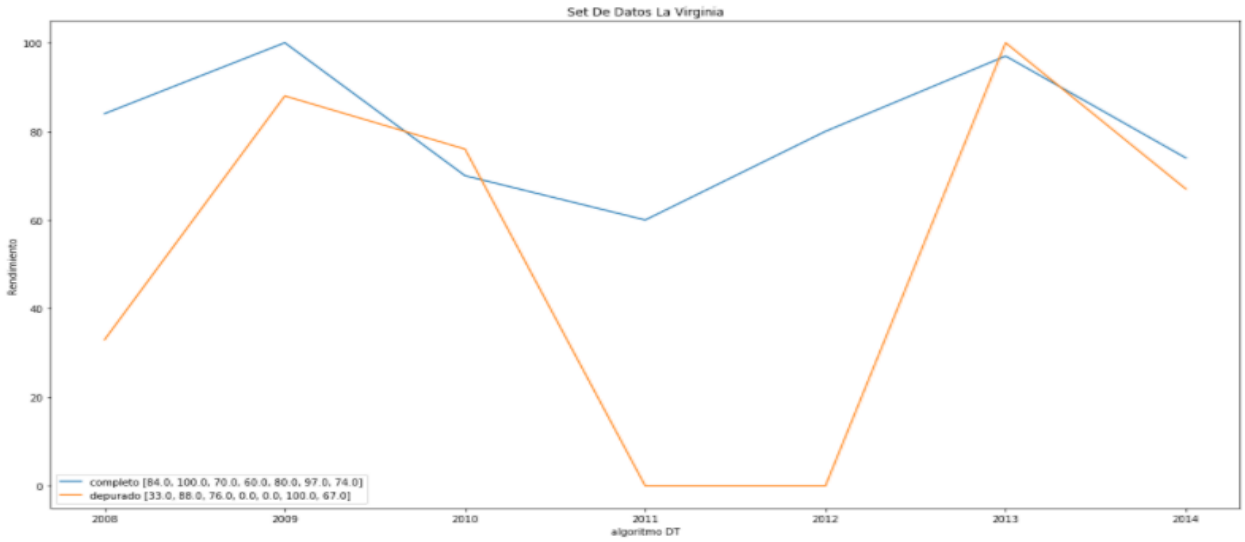


Ilustración 129. Capacidad de Predicción RF La Virginia.

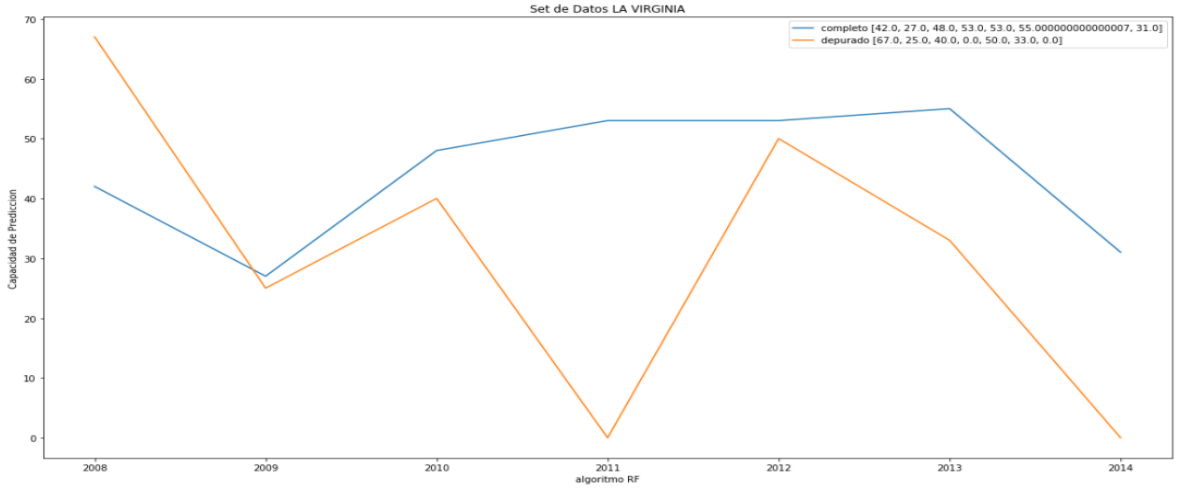


Tabla 79. Capacidad Predictiva La Virginia datos Completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	84	100	70	60	80	97	74
RF	42	27	48	53	53	55	31

Tabla 80. Capacidad Predictiva La Virginia datos Depurados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	33	88	76	0	0	100	67
RF	67	25	40	0	50	33	0

Ilustración 130. Capacidad de Predicción DT Santa Rosa.

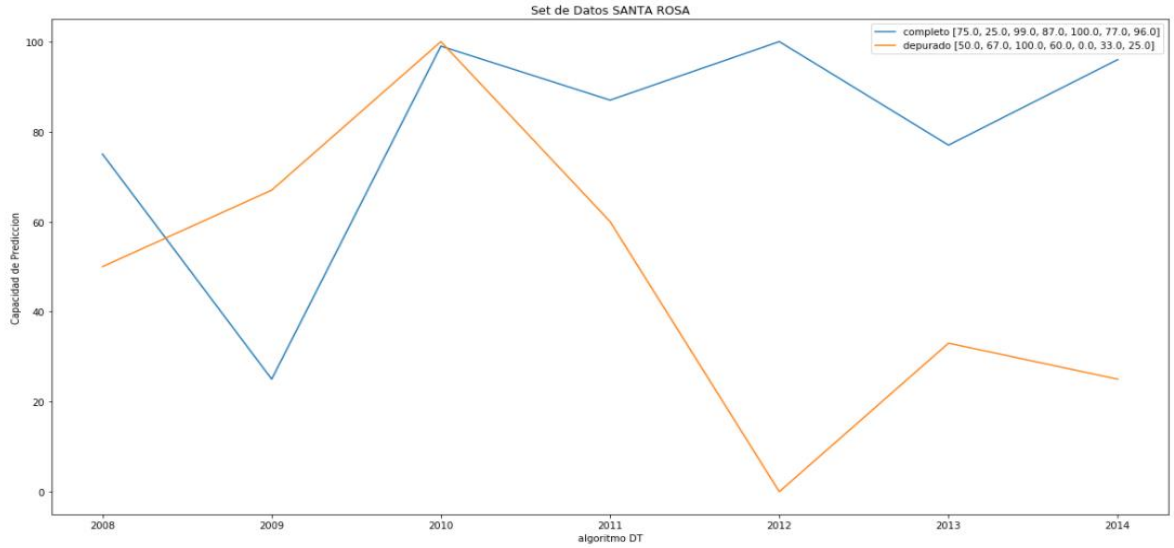


Ilustración 131. Capacidad de Predicción RF Santa Rosa.

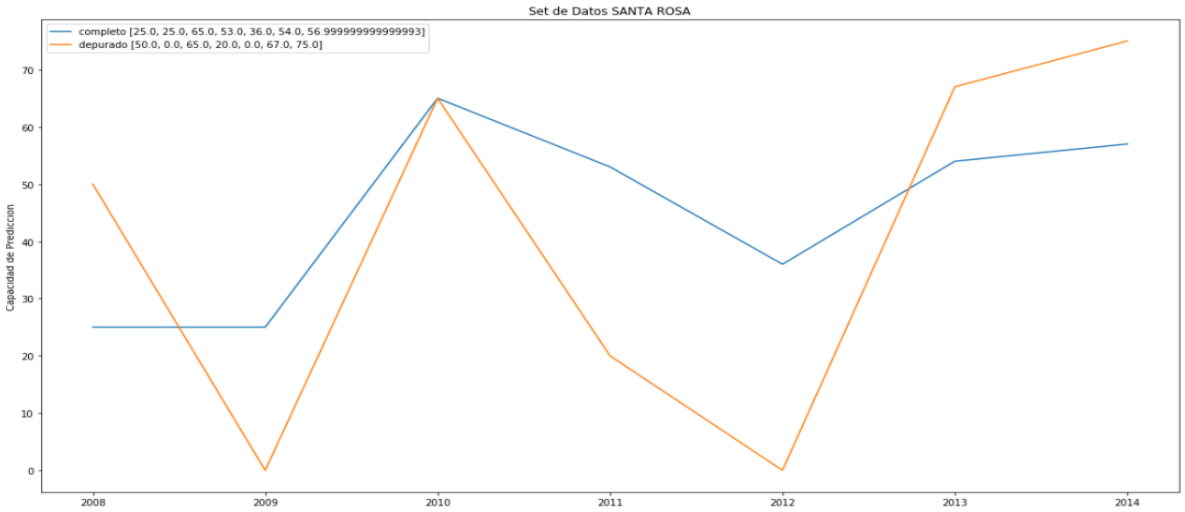


Tabla 81. Capacidad Predictiva Santa Rosa datos Completos.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	75	25	99	87	100	77	96
RF	25	25	65	53	36	54	57

Tabla 82. Capacidad Predictiva Santa Rosa datos Depurados.

ALGORITMO	2008	2009	2010	2011	2012	2013	2014
DT	50	67	100	60	0	33	25
RF	50	0	65	20	0	67	75

Tabla 83. Resumen Aprendizaje Anual DT vs Ciudad Datos Completos.

CIUDAD	2008	2009	2010	2011	2012	2013	2014
SANTA ROSA	75	25	99	87	100	77	96
LA VIRGINIA	84	100	70	60	80	97	74
DOSQUEBRADAS	93	88	97	97	94	85	92
PEREIRA	97	97	97	95	98	97	92

Tabla 84. Resumen Aprendizaje Anual DT vs Ciudad Datos Depurados.

CIUDAD	2008	2009	2010	2011	2012	2013	2014
SANTA ROSA	50	67	100	60	0	33	25
LA VIRGINIA	33	88	76	0	0	100	67
DOSQUEBRADAS	80	82	83	33	0	50	50
PEREIRA	92	96	97	86	75	79	94

La evaluación del nivel de predicción utilizando el algoritmo random forest, si se vio afectada directamente por la cantidad de datos de este estudio como se planteó en la hipótesis, por lo cual la máquina que favorece en mayor proporción el entendimiento del set de datos que se tiene para este análisis, es el árbol de decisiones o Decision Tree, dando un nivel de aprendizaje bastante bueno en el caso de que se tenga una cantidad de registros mínimos para el análisis, por esto se aprecia como varía el porcentaje de predicción del algoritmo cuando se usa el set de datos de todos los casos contra el set de datos que solo usa los casos confirmados, de acuerdo a esto el municipio de Pereira es el único de la muestra que tendría la capacidad de predicción en la línea de tiempo entre 2008 a 2014 de conseguir un acierto de predicción mínimo del 75% de los casos usando la muestra de datos confirmados, para los otros municipios para alcanzar una mejor opción en la predicción es viable trabajar con todos los casos con que se cuenten.

6. CONCLUSIONES

Depurar la información para construir un set de datos interesante para un estudio, es una de las partes fundamentales que permite comprender el comportamiento de un fenómeno, esto puede ser un paso lento e incómodo, pero una parte fundamental del machine learning es contar con unos registros tan bien estructurados y organizados que la parte del análisis podría aportar información importante y concluyente que no se había considerado en el panorama del estudio inicial, permitiendo explicar mejor aún el fenómeno en cuestión.

Un análisis exploratorio de la información permite divisar de manera inicial el comportamiento de las variables que componen un caso de estudio, al tener una investigación descriptiva preliminar. Si consecuentemente se realiza un estudio profundo de la relación entre los componentes de los datos que permita conocer si existen variables que son dependientes de otras o por el contrario son independientes del modelo, mediante tablas o gráficos explicativos que permiten interpretar la información de manera menos abstracta, es posible entender la variabilidad de un fenómeno y que tipo de características se deben incluir o descartar para obtener un mayor beneficio de la investigación que se está realizando.

La fundamentación teórica de esta tesis para estudiar el fenómeno que se quiere describir, fue clave para tomar un punto de partida en el análisis de machine learning, los estudios proponen diferentes tipos de estrategias para abordar el análisis de casos en vectores epidemiológicos, que permitieron tomar diferentes puntos de vista antes de

iniciar la fase de análisis, con lo cual se realizó en este trabajo un análisis con diferentes máquinas de aprendizaje descartando progresivamente las que no permitieron describir el modelo de manera satisfactoria.

Cuando un conjunto de datos contiene en su mayoría datos categóricos en sus clases como el set de datos de este trabajo, es conveniente utilizar clasificadores de tipo árbol, dado que estos regularmente ofrecen un buen nivel de clasificación de los casos, además cuando los valores que tiene en la mayoría de las clases son múltiples, esto repercute directamente en la complejidad del modelo evaluado haciendo que algoritmos como los estudiados en este trabajo, no generen una buena respuesta del clasificador.

Se determinó que si existe una relación entre las variables sociales, económicas y espaciales con el virus dengue transmitidas por vectores epidemiológicos para el set de datos del estudio de este trabajo, donde la condición principal que aprovecha el vector para su ataque en cada municipio es la comuna donde viven los habitantes afectados pero especialmente la cabecera municipal con una incidencia del 75.5%, adicionalmente su condición social representada por el estrato donde residen que son las personas de bajos recursos las de mayor afectación en los estratos 1,2 y 3 que representan el 90% de los casos y donde los de mayor nivel de vulnerabilidad son el estrato 1 que entrega el 50% de los casos totales, otro factor que muestra incidencia de los casos de dengue es la ocupación que desarrolla la persona donde el perfil trabajador no calificado aporta el 90% y el tipo de seguridad social con que cuentan los contagiados que aportan el 80% de los casos, existe otra variable como la edad donde indica que la mayoría de los

afectados no supera los 40 años de edad donde se ubican el 60% de los infectados, dando este estudio como aporte a futuros trabajos que busquen profundizar las variables que se relacionan directamente con el virus dengue.

Un clasificador debe mantener equilibrio tanto en su fase de entrenamiento, como en su fase de verificación, porque esto permite determinar la calidad del trabajo que desempeña, como por ejemplo el análisis de la información de este trabajo con el random forest que parecía ser una excelente alternativa para el análisis del set de datos, porque en su fase de aprendizaje su respuesta fue muy buena, pero al momento de validar su nivel de predicción daba una clasificación baja, por eso se entiende que el algoritmo estaba trabajando sobre ajustado al modelo de datos, lo que obligo a ser descartado del análisis final, aunque el random forest cuenta con un funcionamiento similar al de los árboles de decisión y en varios casos son superiores en muchos aspectos que los arboles de decisión, para este estudio no logró vencerlo y además se vio afectado fuertemente en el momento que la cantidad de datos disminuyó cuando se puso a prueba en el estudio por municipios individualizados.

Utilizar el set de datos completos y el set de datos de casos confirmados de alguna manera fue provechoso, porque esto permitió comparar consecutivamente como se comportaba en fenómeno en la población para este tipo de emergencias, y visualizar que la tendencia de varios tipo de set de datos contaba con algo de similitud aunque la diferencia fuere muy alta en la cantidad de datos, por eso fue necesario en la parte final de análisis solo usar los cuatro municipios que dieron mayor aporte de casos a la

muestra, para tratar de entender el comportamiento de la variabilidad de los casos por municipio en la muestra.

En este estudio se integraron elementos de análisis como el estrato social de las personas infectadas, la ocupación, la seguridad social, el tipo de zona donde habita el afectado y la etnia entre otras, con la que se lograron evaluar el grado de importancia de estas variables en relación a la tasa de infectados para proponer un elemento de distinción en cuanto a indicios sobre inversión en las políticas locales para tratar de disminuir el nivel de impacto de los mosquitos en las zonas más vulnerables ante un eventual ataque de vectores.

Al terminar todo el análisis de las relaciones de las variables en los datos pertenecientes a la encuesta sobre dengue de Risaralda entre los años 2008 a 2014, y evaluar la información con cada una de las máquinas de aprendizaje propuestas para identificar cuál de estas se ajusta mejor al caso de estudio, Se logró obtener un modelo que permite estudiar la incidencia del virus generando un buen margen de predicción en los casos de dengue en Pereira, Dosquebradas, Santa Rosa y La Virginia para el set de datos completo, así como una buena predicción para Pereira en el set de datos de casos confirmados o depurados.

Los aportes de este estudio pueden ser útiles para que las personas, equipos o entidades encargadas de tomar las decisiones relacionadas con las inversiones en las políticas públicas y sociales referentes a la salud como por ejemplo gobiernos locales, regionales

o nacionales puedan crear modernas iniciativas desde nuevas perspectivas, que permitan fortalecer las campañas de control y erradicación de vectores epidemiológicos.

7. RECOMENDACIONES

Los datos de este estudio contaron con una la línea de tiempo muy buena de 7 años, aunque la información contenida en esa encuesta se encontró que el 87.93% de la información estaba mal registrada o incompleta, es de gran importancia obtener en lo posible el mejor set de datos con la mayor cantidad de información correctamente diligenciada por los organismos de control correspondiente, que tienen en su poder los registros de los casos de estudio, aunque la línea de tiempo sea más corta.

El fenómeno de los casos de dengue para los datos de este estudio, mostro efectivamente la existencia de la relación de las variables objetivos de este trabajo, sin embargo, si se contara con los datos como por ejemplo las condiciones sanitarias de las viviendas o barrios de las personas que contrajeron el virus y la zona donde ejercen su ocupación, es posible que se explicara de manera más puntual por que los vectores aumentan su capacidad de infección en una zona determinada.

8. REFERENCIAS

Aalst, W. M. (2016). *Process Mining: Data Science in Action.* Springer.

Abizanda, S. S. (2001). Epidemiología y fisiopatología de la infección perinatal de transmisión vertical. *Sepsis de transmisión vertical XVIII Congreso Español de Medicina Perinatal*, 1-7.

Aguilera-Pesantes, D. R.-P. (2017). Discovering key residues of dengue virus NS2b-NS3-protease: New binding sites for antiviral inhibitors design. . *Biochemical and Biophysical Research Communications.*

Alpaydin, E. (2010). *Introduction to Machine Learning, Second Edition.* Massachusetts Institue: MIT.

Añez G, B. R. (2003). *Impacto económico del dengue y del dengue hemorrágico en el Estado de zulia.* zulis.

Arrieta, G. C.-C.-P. (2017). Zika virus disease, microcephaly and Guillain-Barré syndrome in Colombia: epidemiological situation during 21 months of the Zika virus outbreak, 2015–2017. *Archives of Public Health*, 75(1), 65.

Arshan Nasir, G. C. (2017). Identification of capsid/coat related protein folds and their utility for virus classification. *Frontiers in Microbiology*.

Boshell J, G. H. (1986). Dengue en Colombia. *Biomédica*.

Breiman, L. (2001). Random Forests. *Machine Learning vol 45*, 5-32.

Caicedo-Torres, W. M.-G.-C.-A.-H. (2017). Kernel-Based Machine Learning Models for the Prediction of Dengue and Chikungunya Morbidity in Colombia. . *In Colombian Conference on Computing. Springer, Cham*, 472-484.

Calisher, C. H. (2003). Taxonomy of the virus family Flaviviridae. *. Advances in virus research*, 1-19.

Calisher, C. H. (2003). Taxonomy of the virus family Flaviviridae. *Adv Virus Res. 59*, 1-19.

Carbonell, J. (1990). Paradigms for Machine Learning,. 1-9.

CRUZ ROJA, C. I. (13 de Febrero de 2017). *Colombia: consecuencias humanitarias del conflicto armado en Colombia.* Obtenido de https://www.icrc.org/spa/resources/documents/report/colombia-report-intro-220410.htm

de Almeida Marques-Toledo, C. D. (2017). Dengue prediction by the web: tweets are a useful tool for estimating and forecasting dengue at country and city level. *PLoS neglected tropical diseases*, 11(7), e0005729.

de la Fuente, H. D. (11 de Febrero de 2017). *Cepal.* Obtenido de POLITICAS AMBIENTALES Y DESARROLLO SUSTENTABLE: http://www.cepal.org/publicaciones/xml/6/4496/duran.htm

Demirov, D. G. (2004). Retrovirus budding. *Virus research*, 106(2), 87-102.

Ding, F. F. (2017). Mapping the spatial distribution of Aedes aegypti and Aedes albopictus. . *Acta Tropica*.

Dirección de Epidemiología, A. (2016). *Epidemiología de las Arbovirosis.* Buenos Aires.

Donald S. Shepard, *. L. (2011). Economic impact of dengue illness in the Americas. *Am J Trop Med Hyg*.

Duschinka RD Guedes, M. H.-S. (2016). Zika virus replication in the mosquito Culex quinquefasciatus in Brazil. *Emerging Microbes & Infections*, 1-11.

Endy, T. S. (2010). *DengueVirus:Past,Present and Future, in Frontiers in Dengue Virus Research, K.A. Hanley and S.C. Weaver, Editors.* Norfolk, UK: Caister Academic Press.

Ewout W. Steyerberg, M. J. (1999). Stepwise Selection in Small Data Sets, A Simulation Study of Bias in Logistic Regression Analysis. *Journal of clinical epidemiology*, 935-942.

Fathima, A. S. (2011). A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. . *IJCSI International Journal of Computer Science Issues*, 8(6), 322-328.

Fauran, P. (1996). Prediction and prevention of dengue epidemics. *Bulletin de la Societe de pathologie exotique*, 89(2), 123-6.

Felgaer, P. (2004). Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción. *Reportes Técnicos en Ingeniería del Software*, 6(2), 64-69.

Fernández, I. (1999). *Biología y control del Aedes aegypti: manualde operaciones.* Nuevo Leon: Universidad Autónoma de Nuevo Leon.

Galit Shmueli, P. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R.* Wiley.

Galvis, A. G. (1982). *Historia de la fiebre amarilla en Colombia.* Bogotá: Instituto Nacional de Salud.

Gomes, A. L.-S. (2010). Classification of dengue fever patients based on gene expression data using support vector machines. *PloS one*, 5(6), e11267.

Groot, H. (1980). The reinvasión of Colombia by Aedes aegypti: Aspects to remember. *Am J Trop Med Hyg*.

Grupo de vigilancia y control, d. e. (13 de Agosto de 2010). *MINISTERIO DE SALUD - SIVIGILA.* Obtenido de PROTOCOLO DE VIGILANCIA Y CONTROL DE DENGUE: https://www.minsalud.gov.co/comunicadosPrensa/Documents/DENGUE.pdf

Guo, P. L. (2017). Developing a dengue forecast model using machine learning: A case study in China. *Developing a dengue forecast model using machine learning: A case study in China. PLOS Neglected Tropical Diseases*, 11(10), e0005973.

H Groot, H. V. (1976). Situación epidemiológica del dengue en Colombia. *Boletín Epidemiológico Nacional*.

Halstead, S. (2002). Dengue hemorrhagic fever: two infections antibody dependent enhanment, a brief history and personal memoir. *Revista Cubana de Medicina Tropical* .

Higiene, U. I. (2006). *Bacteriología y Virología Médica.* Montevideo: Oficina del libro FEFMUR.

Hii, Y. L. (2012). Forecast of dengue incidence using temperature and rainfall. *PLoS neglected tropical diseases*, 6(11), e1908.

I. Witten, E. F. (2000). *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations. .* Morgan Kaufmann.

IGAC. (12 de Febrero de 2002). *Instituto Geografico Agustín Codazzy.* Obtenido de Regiones Naturales De Colombia: http://www2.igac.gov.co/ninos/contenidos/detalle_cont_mapas_flash.jsp?varImagen=/ninos/UserFiles/Image/Mapas/REGIONES%20NATURALES.jpg

IGAC. (3 de Febrero de 2017). *Notas Geográficas.* Obtenido de Cuál es la divisón político administraiva actual de nuestro país: http://www2.igac.gov.co/ninos/faqs_user/faqs.jsp?id_categoria=2

J.G RIGAU PEREZ, G. B. (1999). An Evaluation of Modified case definitions for the detections of dengue hemorrhagic fever. *Association of Epidemiologits. PR Healths Sci*.

John Lednicky, V. M. (2016). Mayaro Virus in Child with Acute Febrile Illness, Haiti. *Emerging Infectious Diseases*, 11-22.

Joyanes Aguilar Luis, L. R. (1996). *Fundamentos de Programación.* Madrid: McGraw-Hill.

JS Zuluaga, V. O. (2002). *Nivelación y actualización sobre taxonomía, biología y ecología de Aedes aegypti.* Bogota: Ministerio de Salud, Corporación para Investigaciones Biológicas, Instituto Nacional de Salud.

Kesorn, K. O. (2015). Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. *PloS one*, 10(5), e0125049.

Khan, S. U. (2016). Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). *Biomedical optics express*, 7(6), 2249-2256.

Kok, J. V. (1984). *Construction of plasmid cloning vectors for lactic streptococci which also replicate in Bacillus subtilis and Escherichia coli. .* Applied and Environmental Microbiology.

Krystosik, A. R. (2016). CHIKUNGUNYA, DENGUE, AND ZIKA IN CALI, COLOMBIA: EPIDEMIOLOGICAL AND GEOSPATIAL ANALYSES (Doctoral dissertation, Kent State University).

L Morier, C. E. (2000). Comportamiento biológico de 3 cepas del virus dengue-2 en 2 líneas celulares de mosquitos. *Revista Cubana de Medicina Tropical*.

La Scola, B. D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, 455(7209), 100.

La Scola, B. Z. (2003). Gene-sequence-based criteria for species definition in bacteriology: the Bartonella paradigm. *Trends in microbiology*, 11(7), 318-321.

Lee, V. J. (2009). Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore. . *Tropical Medicine & International Health.*, 14(9), 1154-1159.

Lowe, R. C. (2016). Evaluating probabilistic dengue risk forecasts from a prototype early warning system for Brazil. *Elife*, 5, e11285.

Marino Andres. Alvarez-Meza, A. O.-G.-D. (2017). Kernel-Based Relevance Analysis with Enhanced Interpretability for Detection of Brain Activity Patterns. *Frontiers in neuroscience*, 550.

Martínez y Londoño. (11 de Febrero de 2017). *EL MEDIO AMBIENTE, OTRA VÍCTIMA DEL CONFLICTO ARMADO.* Obtenido de http://ridum.umanizales.edu.co:8080/xmlui/bitstream/handle/6789/2027/Trabajo%20de%20Grado%20Ledy%20Johana%20Martinez%20y%20Maria%20Consuelo%20Londo%C3%B1o%20Holguin.pdf?sequence=1

Ministerio de Ambiente, y. D. (10 de Diciembre de 2012). *Ministerio de Salud y la Protección Social.* Obtenido de DIAGNOSTICO NACIONAL DE SALUD AMBIENTAL: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/INEC/IGUB/Diagnostico%20de%20salud%20Ambiental%20compilado.pdf

Ministerio de Salud, S. d. (07 de 02 de 2018). *Ministerio de Salud*. Obtenido de https://www.minsalud.gov.co/salud/Paginas/SIVIGILA.aspx

Ministerio de Salud, y. P. (15 de Enero de 2017). *Geografía y Salud en Colombia.* Obtenido de ATLAS DE LA SALUD: https://www.minsalud.gov.co/estadisticas/Documents/documento/27-46.pdf

Miroslav Kubat, I. B. (1988). A Review of Machine Learning Methods. *John.Wiley and Sons Ltd*, 3-70.

Morales, A. (1991). El Aedes aegypti en Colombia, historia e importancia en salud pública. *Biomédica*.

Moulton, S. L. (2016). State-of-the-art monitoring in treatment of dengue shock syndrome: a case series. *Journal of medical case reports.*, 10(1), 233.

Mustafa, M. S. (2015). Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control. *Medical Journal Armed Forces India.*, 67-70.

Nagaram, P. P. (2017). Clinical and laboratory profile and outcome of dengue cases among children attending a tertiary care hospital of South India. *International Journal of Contemporary Pediatrics*.

Neuman, K. C. (2008). Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature methods*, 5(6), 491.

Okun, O. (2010). *Applications of Supervised and Unsupervised Ensemble Methods (Studies in Computational Intelligence.* Springer.

OMS. (05 de Febrero de 2017). *Campañas mundiales de salud pública de la OMS.* Obtenido de Información sobre las enfermedades transmitidas por vectores : http://www.who.int/campaigns/world-health-day/2014/vector-borne-diseases/es/

OMS. (6 de Enero de 2017). *Enfermedades transmitidas por vectores.* Obtenido de Centro de prensa: http://www.who.int/mediacentre/factsheets/fs387/es/

OMS. (2017). *Enfermedades transmitidas por vectores.* Ginebra.

OMS, W. H. (1997). *Dengue Haemorrhagic fever: Diagnosis, Treatment,Prevention and Control.* Geneva.

OPS. (1960). *Guía de informes de la Campaña de Erradicación del Aedes aegypti en las Américas.* Washington, D.C.

OPS. (8 de Septiembre de 2010). Obtenido de Organización Panamericana de la Salud: http://www1.paho.org/hq/dmdocuments/2010/alerta_epi_2010_08_septiembre_brote_dengue_corregido.pdf

OPS, M. O. (27 de 12 de 2017). *OPS OMS | Atlas y Mapas Interactivos de Emergencias en Salud*. Obtenido de http://www.paho.org/hq/index.php?option=com_content&view=article&id=13224

%3A-interactive-atlas-and-maps&catid=3889%3Aaro-contents&Itemid=42337&lang=es

Padilla, J. C. (2012). Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia. *Guías de Impresión Ltda*.

Patching, H. M. (2015). A supervised learning process to validate online disease reports for use in predictive models. . *Big data*, 3(4), 230-237.

Pearson, H. (2008). Virophage'suggests viruses are alive.

Pei, J. H. (2011). *Data mining: concepts and techniques.* Morgan Kaufmann.

Quyen NTH1, K. D. (2017). Chikungunya and Zika Virus Cases Detected against a Backdrop of Endemic Dengue Transmission in Vietnam. *The American journal of tropical medicine and hygiene*, 146-150.

Radisic, M. P.-N. (2004). Functional assembly of engineered myocardium by electrical stimulation of cardiac myocytes cultured on scaffolds. *Proceedings of the National Academy of Sciences*, 01(52), 18129-18134.

Reyes-Villanueva, F. (1990). El dengue: Bionomía del vector, transmisión y opciones para su control en Mexico. *Ciencias*.

Richard Sutton, A. B. (1998). *Reinforcement learning: An introduction.* MIT PRESS.

Rodríguez, R. D. (2017). Revisión de literatura de los efectos de cristaloides y/o coloides en el shock hemorrágico.

Rothman, L. (2004). Dengue: defining prolective versus pathologic immunity. *The Journal of Clinical Investigation*.

S. Zientara, D. V.-P. (2015). Evolución reciente de las principales enfermedades transmitidas por vectores Parte I: Panorámica. *Revista científica y técnica Vol. 34 (1)*, 37-39.

Sánchez, L. S. (2011). *CARACTERIZACIÓN ESTRUCTURAL DE FILAMENTOS Y ESTRUCTURAS MEMBRANOSAS INDUCIDAS POR LA LIBERACIÓN DEL VIRUS BUNYAMWERA EN CÉLULAS DE MAMÍFERO.* Madrid.

Shawe-Taylor, N. J. (2007). *Further Reading: Chapter 1.*

Shelley Derksen, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. . *British Journal of Mathematical and Statistical Psychology*, 265-282.

Shi, Y. L. (2016). Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore. *Environmental health perspectives*, 124(9), 1369.

Simon, D. a. (1988). Construction of a vector plasmid family and its use for molecular cloning in Streptococcus lactis. *Biochimie 70*, 559-566.

Soria Segarra Carmen, D. G. (2009). Contribución de Ecuador a la utilización de la clasificación de dengue de la OMS. *Revista Cubana de Medicina Tropical*.

Spratt, H. J. (2013). A structured approach to predictive modeling of a two-class problem using multidimensional data sets. *Methods*, 61(1), 73-85.

Thompson, B. (1995). Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply here: A Guidelines Editorial. *Educational and Psychological Measurement*, 525-534.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, 212-222.

Torres, E. M. (2008). Dengue. *Estudos Avançados vol.22 no.64*, 33-52.

Urbina, E. W. (2017). Presentaciones atípicas de la infección por el virus del dengue: Una Revisión de la Literatura. *Revista Médica de Trujillo*.

Valdes, M. G. (2012). Estudios sobre dengue experiencias y perspectivas. . *e-libro, Corp*, 61 610.

Viana, D. V. (2013). The ocurrence of dengue and weather changes in Brazil: a systematic review. *Revista Brasileira de Epidemiologia*, 16(2), 240-256.

Waggoner, J. J. (2016). Viremia and clinical presentation in Nicaraguan patients infected with Zika virus, chikungunya virus, and dengue virus. *Clinical Infectious Diseases*, 1584-1590.

Wiener, A. L. (2002). Classification and Regression byrandomForest. *R news*, 18-22.